Aalto University
School of Science
Master's Programme in Cloud Computing Services (Specialisation in track ICT innovation)

Mats Mulder

Multiperspectivity in online news: An analysis of how reading behaviour is affected by viewpoint diverse news recommendations and how they are presented

Master's Thesis

Espoo, September 2, 2020

Supervisor: Dr. ir. J.A Pouwelse, Associate Professor (TU Delft)
Dr. N. Tintarev, Assistant professor (TU Delft)
Dr. C. Lof, Assistant professor (TU Delft)
Dr. Antti Ylä-Jääski, Professor (Aalto University)
Thesis advisor(s): Dr. O. Inel, Postdoctoral Researcher (TU Delft)
J.E.G. Oosterman, MSc (Blendle B.V.)

AALTO UNIVERSITY
School of Science
Master's Programme in Cloud Computing Services (Specialisation in track ICT innovation)

ABSTRACT OF THE MASTER´S THESIS

Author: Mats Mulder

Title of the thesis: Multiperspectivity in online news: An analysis of how reading behaviour is affected by viewpoint diverse news recommendations and how they are presented

Number of pages: 133+7

Date: 02-09-2020

Major: SCI3081

Supervisor: Dr. ir. J.A Pouwelse (TU Delft), Dr. N. Tintarev (TU Delft), Dr. C. Lofi (TU Delft), Dr. Antti Ylä-Jääski (Aalto University)

Thesis advisors: Dr. O. Inel (TU Delft), J.E.G. Oosterman (Blendle B.V.)

Previous research on diversity in recommender systems define diversity as the opposite of similarity and propose methods that are based on topic diversity. Diversity in news media, however, is understood as multiperspectivity and scholars generally agree that fostering diversity is the key responsibility of the press in a democratic society. Therefore, a novel viewpoint diversification method was developed, based on the reranking of recommendation lists within the topic using framing aspects. Among other results, an offline evaluation indicated that the proposed method is capable of enhancing the viewpoint diversity of recommendation lists according to a metric from literature. However, to truly enable multiperspectivity in automatic online news environments, users should also be willing to consume viewpoint diverse news recommendation. Therefore, an online study was conducted, assessing how viewpoint diverse recommendations and their presentation characteristics affect the reading behaviour of Blendle users. During a two-week experiment, two groups of 1038 users were presented a set of three recommendations below the content of two articles every day. Thereby, one group received recommendations based on relevance to the original article, while the other group received viewpoint diverse recommendations. Three implicit and one explicit measure of the reading behaviour were analysed. Additionally, the influence of the presentation characteristics of the recommendation on the reading behaviour was analysed. Generally, no major differences were found in the reading behaviour of both user groups. Only the results of the click-through rate calculated per recommendation set indicated a significant difference of 6.5% to the advantage of the baseline users. For the other measures of the reading behaviour, no significant differences were found between the baseline and diverse users. However, the results do show that multiple presentation characteristics have a significant influence on the reading behaviour. Therefore, these results suggest that future research on how recommendation can be presented is just as important as novel viewpoint diversification methods to truly achieve multiperspectivity in automated online news environments.

Keywords: Recommender Systems, Viewpoint Diversity, Online News Media

Publishing language: English

Multiperspectivity in online news

An analysis of how reading behaviour is affected by viewpoint diverse news recommendations and how they are presented

Mats Mulder

Technische Universiteit Delft

TU Delft

MULTIPERSPECTIVITY IN ONLINE NEWS

AN ANALYSIS OF HOW READING BEHAVIOUR IS AFFECTED BY VIEWPOINT DIVERSE NEWS RECOMMENDATIONS AND HOW THEY ARE PRESENTED

by

Mats MULDER

in partial fulfillment of the requirements to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Wednesday September 2, 2020 at 10:30 AM.

Student number:	4288378	
Project duration:	February 28, 2020 – September 2, 2020	
Thesis committee:	Dr. ir. J.A Pouwelse,	TU Delft
	Dr. N. Tintarev,	TU Delft
	Dr. C. Lofi,	TU Delft

An electronic version of this thesis is available at http://repository.tudelft.nl/.

TU Delft

ABSTRACT

Previous research on diversity in recommender systems define diversity as the opposite of similarity and propose methods that are based on topic diversity. Diversity in news media, however, is understood as multiperspectivity and scholars generally agree that fostering diversity is the key responsibility of the press in a democratic society. Therefore, a novel viewpoint diversification method was developed, based on the reranking of recommendation lists within the topic using framing aspects. Among other results, an offline evaluation indicated that the proposed method is capable of enhancing the viewpoint diversity of recommendation lists according to a metric from literature. However, to truly enable multiperspectivity in automatic online news environments, users should also be willing to consume viewpoint diverse news recommendation. Therefore, an online study was conducted, assessing how viewpoint diverse recommendations and their presentation characteristics affect the reading behaviour of Blendle users. During a two-week experiment, two groups of 1038 users were presented a set of three recommendations below the content of two articles every day. Thereby, one group received recommendations based on relevance to the original article, while the other group received viewpoint diverse recommendations. Three implicit and one explicit measure of the reading behaviour were analysed. Additionally, the influence of the presentation characteristics of the recommendation on the reading behaviour was analysed. Generally, no major differences were found in the reading behaviour of both user groups. Only the results of the click-through rate calculated per recommendation set indicated a significant difference of 6.5% to the advantage of the baseline users. For the other measures of the reading behaviour, no significant differences were found between the baseline and diverse users. However, the results do show that multiple presentation characteristics have a significant influence on the reading behaviour. Therefore, these results suggest that future research on how recommendation can be presented is just as important as novel viewpoint diversification methods to truly achieve multiperspectivity in automated online news environments.

Contents

List of Figures

List of Tables

1

INTRODUCTION

1.1. INTRODUCTION TO THE PROBLEM

Due to the expansive growth of information accessible on the internet, research to so-called *recommender systems* has already been introduced in the 1980s [70]. Recommender Systems are software tools and techniques providing suggestions for items to be of use to a user. These suggestions are aimed at supporting their users in various decision-making processes, such as what items to buy, what music to listen, or what news to read [68, 72]. In today's world, the majority of widely used internet services provides some form of personalisation using recommender systems. In 2006, Amazon reported that 35% of its sales derived from recommender systems [46]. Netflix reported in 2012 that 75% of the video streams are selected by the users from recommendations lists [2].

Likewise, the online news industry is increasingly using recommender systems to personalise their content. In recent years, this form of news distribution has grown significantly. In 2015, 23% of the interviewees of research by Newman, Levy, and Nielsen reported online media as their main news source and 44% considers digital and traditional sources equally [58]. Consequently, not only traditional news media are increasingly distributing their content online, also digital-born news websites and *news aggregators*, which combine content from various sources in one service, are gaining ground. For example, the share of digital-born news platforms in Australia and Japan is already larger than the share of traditional news media [58]. Additionally, the digital form enables real-time updates, thereby increasing the distribution speed [38].

Scholars generally agree that the key responsibility of the press in a democratic society is to expose of citizens to a high diversity of viewpoints on a particular topic, enabling citizens to act well-informed in their decision making process [51, 79, 63, 21]. However, the capability of current recommender systems to support this responsibility, by providing a high diversity of viewpoints on a particular topic, can be questioned. Recently, this issue became widely known as filter bubbles and echo chambers [61]. The high level of personalisation would lock up people in bubbles of what they already know or think and

1

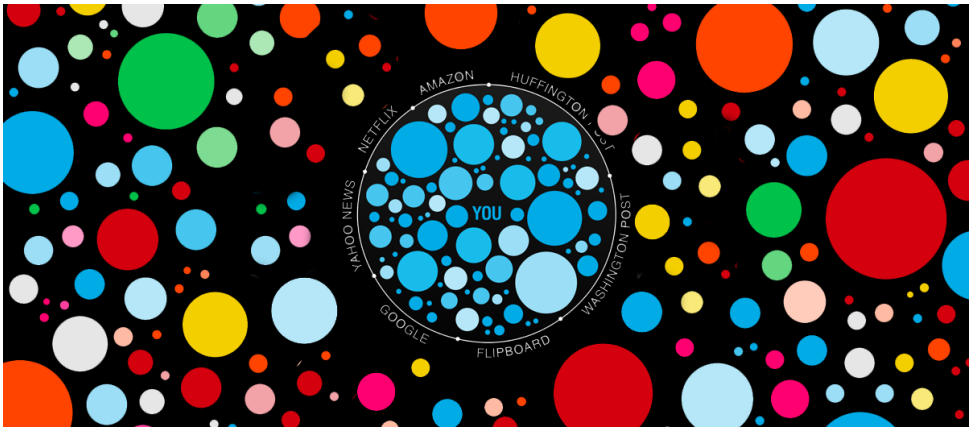


Figure 1.1: Illustration of the potential thread of filter bubbles according to Pariser [61]

'rooms' with only like-minded people [61, 54, 7]. Although the real existence of these phenomena in current recommender systems is under discussion -only minor effects have been found so far [7, 29, 59]-the ability to diversify recommendations based on viewpoint could prevent these issues in future systems. Moreover, viewpoint-aware recommendation systems would be able to preserve the quality-level of a news landscape during the shift to an increasing online industry.

Current diversification methods for recommender systems focus mainly on topic diversity and thus, are not directly applicable in the news domain to increase viewpoint diversity. Therefore, novel diversification methods are needed that are capable of enhancing the diversity of viewpoints in recommendation lists. However, to truly enable multiperspectivity in automatic online news environments, users should also be willing to consume viewpoint diverse news recommendation. Therefore, research should also address how the reading behaviour is affected by viewpoint diverse news recommendations and their presentation characteristics.

1.2. INTRODUCTION TO BLENDLE

Blendle is a Dutch startup founded in 2014. The company offers an online news platform that aggregates articles from more than 150 national and international newspapers and magazines. Although their initial business model was based on payment per article, their current business model includes an 10,- euro monthly subscription. Every day, users receive a recommendation list including 20 articles from national and international titles. Additionally, subscribed users can explore the Blendle archive, including articles from newspapers older than seven days and all articles from magazines, for free. The daily selection can be seen as the most important aspect of the Blendle product and is a combined effort of both the Blendle editorial team and a recommender system.

1.2.1. BLENDLE ARTICLES

As earlier mention, Blendle aggregates more than 150 national and international newspapers and magazines. However, an important note to make is that Blendle only includes articles that are originally published on paper. For example, short articles of daily news incidents that are only published online will not be included. Therefore, the term news articles, when used in the study, refers to articles from newspapers and magazines that are originally published on paper.

1.2.2. RECOMMENDER SYSTEM

As described above, the daily recommendation list is composed by both the Blendle editorial team and a recommender system. The editorial part includes around five articles and will be served to all users. The other part will be supplemented by a recommender system, which generates a personalised selection for every user. The recommender systems uses both implicit information, such as the read articles, and explicit information, such as feedback on an article, to learn the user's preferences. The generation of this recommendation list for a certain user can be described by the following five-step pipeline:

1. **Data Enrichment**
   Blendle receives articles from publisher as plain text. At arrival, articles are enriched using several methods based on regular expressions, algorithms and machine learning models. The enrichment varies from simple information, such as author and title, to more complex details, such as entity predictions, complexity, feel and quality. The retrieved information is essential for the sequential steps of the described pipeline and other Blendle systems.

2. **Input Selection**
   Although the Blendle archive comprises around three million articles, the input of the recommender system is already filtered on essential features, such as relevance and quality. For example, articles should have been published recently (depending on the publisher) and not already read by the user. Every day, this list comprises around 500 to 700 articles. The editorial selection is always included for everyone.

3. **Article Ranking**
   During this step, the articles obtain a ranking score for each user. Currently, this ranking is a linear combination of five features and their corresponding weights. Some examples of the features for an article A include:

   - Does a user follow the publisher of A?
   - How many articles of the same topic has the user read?
   - Did the user provide negative feedback on a publisher?

4. **Diversification**
   During this step, the actual selection of 20 articles is composed for every user. First, the editorial selection is added. Afterwards, articles are added using a *Maximal Marginal Relevance algorithm* (MMR). In a nutshell, the algorithm adds articles one by one based on a threshold between the relevance of the article for the user

(ranking score) and the dissimilarity with all previously added items [15]. The dissimilarity between two articles is calculated as a linear combination between the number of overlapping entities and a penalty if both articles are assigned to the same cluster based on a hierarchical clustering algorithm. An important remark to make: the MMR does not take the already-added editorial articles into account when calculating the similarity. This is based of the idea that the editorial and algorithmic selection should be independent.

5. **Presentation**
   Finally, the selection of 20 articles is processed to be presented to the user. Two different presentation methods can be distinguished:

   A **Email**
   In the morning, 12 articles of the daily selection will be send to users (who subscribed for the email). All editorial selections are included at the top. Below, the top articles from the MMR are added.

   B **Blendle Platform**
   The Blendle product is available on iOS, Android and web. All platforms follow the same layout to present the daily selection. On top, the editorial selection is presented in a dedicated section. Below, the algorithm selection is presented in groups of articles of overlapping topic or publisher. These groups are formed using a similar method as the MMR.

1.2.3. RESEARCH IN THE CONTEXT OF BLENDLE

To better understand the intention of this study in the context of Blendle, we describe two possible features that could be implemented related to the focus of this thesis.

As earlier mentioned, the current recommender system of Blendle generates one part of the daily news recommendation for users. Concretely, this includes around 15 articles per day. Because of the wide variety of included sources and topics in the Blendle platform, it is not very common that two articles on the same topic are included on the same day. Consequently, it is expected that the potential diversification of viewpoints for the daily list is relatively small. When considering the recommendation over a longer time-frame, however, viewpoint diversity of the recommendation could potentially increase. For example, an article could have a higher probability of being included in the daily recommendations if it represents a different viewpoint on a certain topic compared to an article that was suggested earlier that week.

Figure 1.2: Example of Possible Blendle Feature

Another interesting example includes the recommendations below an article. Currently, a part of the daily recommendation is repeated at the end of every article. This is illustrated in figure 1.2. Instead of recommendation on other topics, this space could also be used for recommendation that have a different viewpoint on the same topic.

Thereby, one can actively suggest users to consider different viewpoints on the same topic and so, prevent issues like the creation of filter bubbles.

Note, however, that these described features are only used to illustrate the potential of the research in the context of Blendle and that the actual implementation of such features in the Blendle product is not a part of the thesis work. The focus of the thesis includes the viewpoint diversification methods that could be used by such features.

1.3. RESEARCH QUESTION

As described in the introduction to the problem, current diversification methods for recommender systems focus mainly on topic diversity and thus, are not directly applicable in the news domain to increase viewpoint diversity. Therefore, novel diversification methods are needed that are capable of enhancing the diversity of viewpoints in recommendation lists. In the end, however, online news readers should also be willing to consume viewpoint diverse news recommendation. Therefore, to enable true multiperspectivity in the online news environment, research should also address how the reading behaviour is affected by viewpoint diverse news recommendations and how they are presented. Therefore, the main research question is defined as follows:

Main RQ: *How is reading behaviour affected by viewpoint diverse news recommendations and how they are presented?*

1.3.1. SUB QUESTIONS

To be able to answer the main research question, multiple sub-question have been defined. The questions are ordered according to the outline of this report:

1. **How is diversity defined in the context of news media? What conceptualisation can be used to diversify news recommendation?**
   As described before, current diversification approaches are not applicable in the news domain. Therefore, this study starts with a literature study on how diversity is defined in the context of news media and what conceptualisation can be used to diversify news recommendations.

2. **What metadata can be related to this conceptualisation and which methods and tools can be used for the extraction of this data?**
   To be able to use this identified conceptualisation in a diversification method, metadata that can be related to the conceptual aspect should be determined. Additionally, suitable methods to extract this metadata need to be chosen.

3. **How can this metadata be combined to a measure for viewpoint diversity that can be used in a recommender system?**
   Thirdly, the extracted metadata needs to be combined to a global diversity measure that assess the dissimilarity of two articles in terms of viewpoint. Afterwards, a suitable diversification algorithm that is based on this measure should be identified.

4. **Is the proposed method capable of increasing the viewpoint diversity of recommendation lists, according to a metric from literature?**
   As final step before the main research question can be addressed, the capability of the proposed diversification method to enhance the viewpoint diversity of new recommendation lists should be assessed, according to a metric from literature.

1.4. SCIENTIFIC CONTRIBUTIONS

In this work, the following scientific contributions are made:

- A literature study is conducted on what definition and conceptualisation of diversity in the context of news media can be used in a viewpoint diversification method
- A data set on suggestions in news article content is composed using a crowdsourcing task. The data set is used to evaluate rule-based methods for suggestion mining.
- A novel viewpoint diversification method based on framing aspects is proposed that is capable of enhancing viewpoint diversity in recommendation lists within the topic using reranking.
- An online evaluation is conducted on how viewpoint diverse recommendations and their presentation characteristics affect the reading behaviour of users.

2

Background and Related Work

2.1. Introduction

First, a literature study is performed to both gain insights in the research domain and evaluate related work. As shortly described in the introduction, current diversification methods are not directly applicable in the news domain. To elaborate more on this issue, this chapter starts with a section on current approaches of diversification in recommender system in section 2.2. Afterwards, section 2.3 investigates how diversity is understood in social sciences and section 2.4 discusses the concept of framing as possible conceptualisation for novel diversification methods. Section 2.5 provides a brief introduction to natural language processing techniques and tools that can be used to extract metadata from news article content that can be used by the diversification method. Finally, the findings of the chapter our discussed in the conclusion in section 2.6.

2.2. Diversity in Recommender Systems

The following section will provide an overview of literature on diversity in recommender system. First an introduction to recommender systems is given. Afterwards, the raise of diversity as a measure for recommender systems is described. Thirdly, the current gap between diversity in recommender systems and diversity in news media is described. Finally, a summary and conclusion are provided.

2.2.1. Introduction to Recommender Systems

Due to the expansive growth of information accessible on the internet, research to so-called *recommender systems* has already been introduced in the 1980s [70]. Recommender Systems are software tools and techniques providing suggestions for items to be of use to a user. These suggestions are aimed at supporting their users in various decision-making processes, such as what items to buy, what music to listen, or what news to read [68, 72]. In today's world, the majority of widely used internet services

provides some form of personalisation using recommender systems. In 2006, Amazon reported that 35% of its sales derived from recommender systems [46]. Netflix reported in 2012 that 75% of the video streams are selected by the users from recommendations [2].

Several different types of recommender systems have been proposed so far. The most commonly known are *collaborative filtering* algorithms, including user-based and item-based collaborative filtering approaches. User-based collaborative filtering constructs recommendation lists using similarities between users. Item-based collaborative filtering focuses on the similarities between items directly. The recommendation list is constructed by comparing items the users liked with all other items. Other types of recommender systems include content-based systems, hybrid approaches or specific prediction generation algorithms, such as support vector machines and k-nearest neighbours methods [43].

2.2.2. DIVERSITY AS A METRIC FOR RECOMMENDER SYSTEMS

Traditionally, research on recommender systems evaluated their method based on accuracy metrics, such as precision, recall and mean absolute error [94]. The focus on accuracy, however, induced a problem which is known as *over-fitting*. Thereby, a model is fitted so strongly to an user that it will be unable to detect any other interests. This problem is also known is overspecialisation [43].

To overcome this issue, there was a need for beyond-accuracy metrics. Thereby, the focus was shifted to a more user-centric evaluation of recommender systems, including diversity [94]. As a result, multiple studies have evaluated the influence of diversification on the user experience. Most of these studies indicate a positive influence of diversification of recommendations on the user experience [94, 40, 90, 51, 19]. Moreover, some studies were able to preserve high-levels of both relevance and diversity, where this is often considered as a trade-off [82, 18].

Because of its positive effect on the user experience and its capability to prevent the over-fitting issue, diversity is currently considered as an important metric for recommender systems. Thereby, diversity is most commonly defined as the opposite of similarity [43]. Consequently, most approaches on diversification in recommender systems focus on topic diversity. For example, Ziegler et al. proposed a method to rerank a list of recommendations based on topic diversity. For that purpose, a new metric called *intra-list similarity* is proposed that captures the similarity of each pair of recommendations [94]. The evaluation of the method using both an user-based and item-based collaborative filtering algorithm shows that, although precision and recall decrease, the user satisfaction with the system increases [94].

2.2.3. DIVERSITY IN THE NEWS DOMAIN

Although there is increasing attention for diversity in recommender systems, these approaches are not directly applicable in the news domain. As described before, diversity in recommender systems is often defined as the opposite of similarity and approaches are based on topic diversity. Diversity in the news domain, however, is an extensively researched concept in social sciences and is often understood as multiperspectivity or a diversity of viewpoints [24]. Thereby, it is argued that the exposure of citizens to a high

diversity of viewpoints on a particular topic is a crucial aspect of any democratic society, enabling citizens to act well-informed in their decision making process. Consequently, viewpoint diversity is considered as a key principle of news quality within any democratic society around the world [51, 79, 63, 21].

To our best knowledge, only one study has proposed a method for viewpoint diversification in recommender systems so far. Tintarev et al. propose a news distance measure for viewpoint diversity and implement this in the *Maximal Marginal Relevance* (MMR) algorithm [82]. Thereby, the distance measure was composed as a weighted combination of article variables, such as emotional tone, article source and linguistic complexity. The weights were optimised using a grid search over all model variables. Afterwards, the diversity measure was implemented in the MMR algorithm, which reranks a list of recommendations based on a linear combination between relevance and diversity [15]. The parameter λ represents the threshold between relevance and diversity. During an offline study, the method was evaluated using the Intra-list Diversity metric using Channels and Topics features. For that purpose, a dataset was created, containing 386 English articles from 17 sources. The results indicate a positive effect of the method on the viewpoint diversity metric.

Although the study indicates a positive effect on the viewpoint diversity, the relation between the implemented feature vector and the conceptualisation of diversity in news media is indistinct. This study aims to propose a novel diversification method by bridging the conceptualisation of viewpoint diversity in the social domain to the computer science domain. Additionally, the method will be focused on Dutch news articles. As a starting point of this process, a literature study on diversity in news media is conducted. This can be found in section 2.3.

2.2.4. SUMMARY

The section can be summarised by the following points:

- Recommender systems are indispensable in the current internet landscape. This increasingly holds for the news domain as well.
- To overcome the over-fitting problem, new metrics including diversity have been introduced in the research domain of recommender systems.
- Most studies on diversification define diversity as the opposite of similarity and propose methods that are based on topic diversity. Diversity in news media, however, is understood as as multiperspectivity or a diversity of viewpoints.
- Therefore, novel methods are needed that are able to bridge viewpoint diversity in the social domain to the computer science domain. As a starting point, the next section will elaborate on diversity in news media.

2.3. DIVERSITY IN NEWS MEDIA

The following section provides on overview of the main insights found in literature on the concept of diversity in news media. The first subsection describes how the concept is commonly understood on social sciences. Afterwards, two general approaches

in evaluating diversity are described: source and content diversity. Finally, a summary and conclusion are provided.

2.3.1. Diversity as a Key measure for News Quality

Different from the definition of diversity in recommender systems, defined as the opposite of similarity, diversity in news media is understood as multiperspectivity or a diversity of viewpoints [24]. In the social domain, diversity is generally seen as a key measure for the news quality [63, 16, 48]. Thereby, most studies follow a paradigm that directly relates diversity in news media to normative standards on the principles of a stable and effective democracy [55, 89]. This paradigm describes diversity in news media as a marketplace of ideas, in which citizens choose from a wide variety of ideas, delivered from a wide range of sources. Thereby, it promotes broader social objectives of democracy, including informed decision-making, cultural pluralism, citizens welfare and a well-functioning democracy. Following this paradigm, scholars generally agree on the key responsibility of the press in fostering diversity [6, 11, 22]. For example, Gans states that, instead of aiming for the unattainable goal of being objective, the press needs to be *multiperspectival* [24].

At the same time, however, the exact aspects and measurement of viewpoint diversity are often disputed and several different approaches exist [55, 5, 11]. Generally two main approaches can be distinguished: source and content diversity.

2.3.2. Source Diversity

Most studies that assess diversity in news media, focus on the diversity of sources [55, 5, 89, 4]. Source diversity can be defined as "a dispersion of the representation of affiliations and status positions of sources used to create a news product" [89]. Here applies: more dispersion results in more diversity.

Measurement of Source Diversity

When measuring source diversity, most methods follow Bennett's indexing theory, which assumes that the inclusion of nonofficial or nonelite sources corresponds to high levels of diversity [10, 5]. Alternatively, Napoli approaches the issue from a policymaker point of view and distinguishes three aspects of source diversity. The diversity of ownership of content or programming, the diversity of ownership of media outlets and in terms of the diversity of the workforce within individual media outlets [55].

The effect of Source Diversity

Besides measuring, some scholars study how source diversity is influenced by certain aspects. Some studies assess the effects of the proximity, the geographical distance between event and source. An analysis by Hackett reveals the increasing reliance of news media on high-level or official sources, such as governments, administrations or foreign officials, when the proximity is large [28]. Likewise, Martin found an inversely proportional relation between the proximity and the number of sources that are covered [47].

Limitations of Source Diversity

While many studies have addressed diversity using source diversity, or even see source diversity as a measure for content diversity, critics agree on its limitations. As Voakes

et al. puts it: "it is entirely possible that a story can contain quotes and information attributed to a wide variety of source types, from unaffiliated nobodies and corporate executives to publicists and protesters. But if all of the attributions revert to the same frame or point of view, then we must question whether diversity is truly in evidence in that story" [89]. On the other hand, one source could provide multiperspectival views on a topic. Thus, although diverse sources are likely to increase the inclusion of diverse interpretations, it is no direct measure [5, 89]. This argument is supported by multiple studies on aspects that influence source diversity. For example, Benson shows how certain aspects, such as power distributions in society, commercial pressure of news media and journalistic norms and practices, significantly influence which sources gain access to the media [11, 5]. Smith et al. indicate that if non-elites are cited, only a a limited number of sources, known as activist organizations, is referenced [76].

2.3.3. CONTENT DIVERSITY

On an abstract level, content diversity is defined by Van Cuilenburg as "heterogeneity of media content in terms of one or more specified characteristics" [83]. Others have conceptualised it more concretely, such as Voakes et al., who understand content diversity as "a dispersion of representation of ideas, perspectives, attributions, opinions, or frames within a news product, and within the context of one particular issue" [89].

Based on the limitations related to source diversity, most scholars agree that diversity can only by achieved by fostering content diversity [48, 55, 16, 24, 6, 89]. As Voakes et al. put it: "Ultimately, it is the content of news that transmits diversity directly to the audience -not the personal and processes by which the news was gathered. The content is what activates, motivates, interests and involves its mass audience. It is our contention, therefore, that a principal dimension of this elusive concept of diversity should be content diversity" [89]. Likewise, Choi argue that content diversity is ultimately an indicator of the quality of news reporting [16].

MEASUREMENT OF CONTENT DIVERSITY

However, scholars that do acknowledge content diversity as key aspect of diversity in news media, use multiple divergent measurement methods. In a recent study, Baden and Springer identified six common approaches in studies assessing content diversity [5]. Within these six approaches, two subdivisions can be recognised. The first three methods focus on the tone or political position represented in the news. The simplest approach assumes high levels of content diversity if nonofficial or nonelite views are present in the debate. Other attempts suppose that viewpoint diversity can be obtained by a variety of tones attached to political issues, or aim to recognize discern political bias in news coverage. All these approaches, however, are confronted with the same major issue as source diversity; diverse political standpoints could relate to identical interpretations of an issue and thus, they provide no direct measure for viewpoint diversity [5]. Moreover, several studies reveal additional drawbacks of these methods. Among other things, scholars have shown how privileged relations between the media and selected political and social groups strongly influence news production [3, 33]. Also, some studies indicated that viewpoints that are already present in news discourse tend to remain and prevent new, marginal views from gaining access [85, 69]. Moreover, Porto argues

that these methods are too uni-dimensional: "citizens need a broader variety of cues in the news media than those resulting from the traditional routine of hearing both sides" [63]. The second three approaches, as described by Baden and Springer, concentrate on the description of political issues. Within these, a first method uses the diversity of language to evaluate content diversity. However, this is again no direct measure, since the same perspective can be described through different language.

The final two approaches use the concept of *frames* to assess content diversity. Framing theory states that every communicative message selectively emphasizes certain aspects of the complex reality [5]. Thereby, frames enable different interpretations of the same issue [73]. Framing has been put forward by many scholars as conceptualisation for enhancing content diversity. For example, Porto proposed a novel *citizen competence model*, which describes the conditions that facilitate or prevent the fulfillment of citizens' civic roles, and a quality standard for news media based on this model [63]. His *interpreting citizen model* assumes that citizens are able to fulfil their civil roles if they are able to interpret political issues in a consistent way. Thereby, it is assumed that a plural news environment guarantees this consistency. The *News Diversity Standard* judges news media on their performance providing such an environment by means of presenting diverse frames. In a more recent study, Baden and Springer describe three aspects of frames that are central to viewpoint diversity's role in democratic media. First, frames create different interpretations of the same issue by selecting some aspects of the complex reality [23]. Secondly, frames are not neutral, but suggest specific evaluations and courses of actions that serve some purpose better than other Entman. This aspect forms the basis for the final argument: frames are often strategically constructed to advocate specific political views and agendas.

Framing, thus, is generally seen as one of the most suitable conceptualisations of content diversity. However, the concept is an extensively researched concept within a variety of research fields, including communication, sociology and psychology [63]. Also, different definitions, conceptualisations and variants are described in literature. Therefore, the next section will elaborate on the concept of framing.

2.3.4. SUMMARY

The section can be summarised by the following points:

- Generally, scholars agree that providing diverse viewpoints is the key responsibility of press. Thereby, it promotes broader social objectives of democracy, including informed decision-making, cultural pluralism, citizens welfare and a well-functioning democracy.
- Content diversification and source diversification methods can be distinguished. However, scholars argue that viewpoint diversity can only by achieved by fostering content diversity.
- One promising approach for content diversification includes the diversity of frames. Framing theory is described in the next section.

2.4. NEWS FRAMING

This section describes literature on framing theory. The first subsection will describe what is understood as framing and how different studies can be distinguished. Afterwards, three main study domains related to framing are discussed. Finally, the section is summarised and concluded in the last subsection.

2.4.1. FRAMING AS A PROCESS

Framing is an extensively researched concept among multiple different domains, including psychology, communication and sociology. Its roots can be found in the latter domain; In 1955, Bateson stated that communication only gets meaning in its context and by the way the message is being constructed [8, 88]. Later, frame theory gained increasing momentum and was generally understood as follows: every communicative message selectively emphasizes certain aspects of complex reality [5]. Therefore, every news article unintentionally comprises some form of framing [5]. Thereby, several aspects such as the choice of certain topics, the inclusion of specific sources, or the choice and placement of words in an article contribute to a particular frame. Additionally, frames are often deliberately used to construct strategic, often political, views on a topic. Consequently, frames enable different interpretations of the same issue [5]. However, this also implies that every frame inevitably deselects other, equally plausible and relevant frames [5].

As a consequence of the wide adaption of framing theory in different research fields, scholars have argued that the concept is referred to with significant inconsistency in literature [17]. To overcome this inconsistency and congregate different interpretations of framing in literature, De Vreese proposed an integrated process model of framing, including *frame locations* and *frame stages* [17]. Frame locations describe the manifestation of frames, whereas frame stages describe how different locations affect each other. Starting at the beginning of the process, research on *frame building* addresses the question how frames in the newsroom affect the manifestation of frames in the news content. For example, which internal factors of an editorial team affect the presence of certain frames in the news. Secondly, research on *frame setting* describes how frames in news content affect society. For example, what attitudinal effects does a certain frame has on its reader? An visual overview of the integrated process model can be found in Figure 2.1.

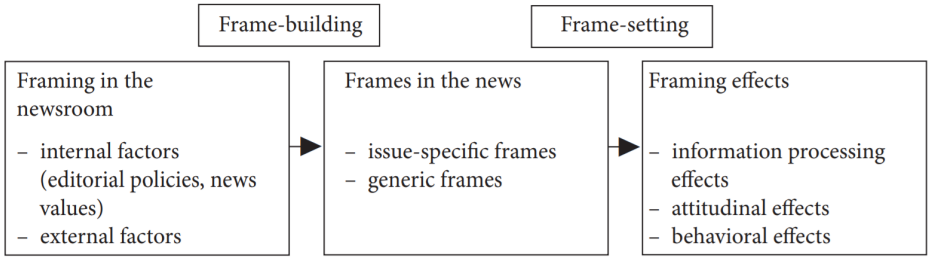


Figure 2.1: an integrated process model of framing [17]

Since the study is interested in the extraction of information related to framing from news articles, the next section will elaborate on the manifestation of frames in text, including definitions of frames in news articles and conceptual differences. Afterwards, a brief introduction to frame building and frame setting are given to provide some more insights on the concept as a whole.

2.4.2. FRAMES IN THE NEWS

First, the concept of frames in the news, including relevant literature, will be discussed. Although the processes of frame building and frame setting are key aspects as well, understanding what constitutes a frame is essential in the light of this study. Therefore, definitions of frames in literature will be a good starting point.

DEFINITIONS

As mentioned before, framing theory has been adopted broadly, including multiple different research domains. Consequently, several definitions can be found in literature. According to Giltin, who performs a study on framing of the news left in the U.S., frames can be defined as "principles of selection, emphasis and presentation composed of little tacit theories about what exists, what happens and what matters" [26]. Almost a decade later, Gamson and Modigliani evaluated media and public opinion on the U.S. nuclear movement [23]. In their study, frames are understood as "interpretative packages" that give meaning to an issue [17]. In this description, frames form the central aspect of the interpretative packages. The definition by Entman, however, is most commonly adopted in literature. This definition states that framing includes the selection of "some aspects of perceived reality and make the more salient in a communicating text, in such a way as to promote a particular definition of a problem, causal interpretation, moral evaluation and/or treatment recommendation for the item described" [20]. Within this definition, the problem describes "what a causal agent is doing with what costs and benefits, the diagnose causes "identify the forces creating the problem", the moral judgements "evaluate causal agents and their effects and the suggested remedies "offer and justify treatments for the problems and predict their likely effects". Additionally, Entman describes how frames can be found at different levels of analysis, including single sentences, paragraphs or articles as a whole. Also, a frame may not necessarily include all four functions as described before [20]. Although this definition is most commonly used, literature does not often conceptualise the definition in a consistent manner [88]. Therefore, less strict "working definitions" have been proposed, such as the definition by De Vreese: "a frame is an emphasis on salience of different aspects of a topic" [17]. However, this definition has been criticised for even further stretching the concept [88].

Therefore, this study will use the definition by Entman [20]. Primarily because of the wide-adaption in related literature, including recent studies. Additionally, the concrete functions of a frame, as described in this definition, are better suited to be translated to computational equivalents.

CONCEPTUAL DIFFERENCES

Considering differences in conceptualisation of frames in literature, one major distinction can be made between *issue-specific* and *generic* frames. Although some studies de-

fine other conceptualisation or do not necessarily acknowledge this distinction, most studies can be divided into these two categories in retrospective.

Issue-specific frames are connected to specific issues or events [17]. Often, they also relate to a specific context or time [88]. Issue-specific frames have the advantage of being detailed about the issue under investigation. However, this also relates to its major drawback; The high level of detail makes it difficult to generalise and compare frames related to different issues. Thereby, it prevents theory building of framing [17]. According to Hertog and McLeod, this has resulted in "researchers finding too easily evidence for what they are looking for" and "the frustrating tendency of generating a unique set of frames for every study" [31]. This statement is supported by a literature analysis of Matthes, who identified 561 different issue-specific frames among 131 articles on framing [49]. Additionally, he found that 78% of the studies focuses on issue-specific frames. A study by Shah et al., can be seen as an example of literature on issue-specific framing. In their study, three frames related to final stages of the Clinton presidency are identified: "Clinton behaviour scandal", "Conservative attack scandal" and "Liberal response scandal" [75].

In contrast to issue-specific frames, generic frames can be associated with different issues. Moreover, they can be identified independent of context or time [17, 88]. Opposite pros and cons apply to generic frames compared to issue-specific frames; While issue-specific details get lost, generic frames allow for comparison across studies. Often, this also implies a standard way of measuring [88]. Critics, however, wonder whether these are indeed frames, or rather "arguments" or "scripts" [84]. Among the 131 articles evaluated by Matthes, 29 generic frames were found [49]. Examples include the "conflict frame", the "issue frame", the "thematic frame", the "attribution of responsibility frame", "strategy frame" and the "economic consequence frame". An example of studies on generic frames includes the identification of frames in media during election campaigns. Adriaansen, Van Praag, and De Vreese indicated that around fifty percent of the television items and newspaper articles included a "strategy frame" during the 2006 Dutch elections [1].

2.4.3. FRAME IDENTIFICATION AND EXTRACTION METHODS

When identifying and annotating frames, different methods in literature can be distinguished. Probably, the main distinction includes the difference between *deductive* and *inductive* approaches. Deductive frame studies define the frames to be identified before the actual analysis. Inductive methods, in contrast, include frame identification as important step, or are even the goal, of the analysis [88]. Most inductive studies use a pre-study on a part of the content to identify the frames [88]. Critics, however, argue that inductive matters rely on too small samples and are therefore, difficult to replicate [17]. In addition to the distinction between inductive and deductive studies, the variables, or *frame devices*, under investigation can be different. Examples include the use of keywords, catchphrases or metaphors in text or frame devices in the form of visual content, such as images. While most studies acknowledge the role of visual content for framing analysis, 83% completely neglects this in their method [49]. Finally, a distinction can be made between studies that have a quantitative or qualitative approach and the use of computer-based assistance [88, 49]. A study by Matthes and Kohring discusses five

categories of frame identification and annotation methods based on these characteristics. Although the categories are not mutual exclusive, it provides a credible overview of related literature. The five categories include the *hermeneutic approach*, the *linguistic approach*, the *manual holistic approach*, the *computer-assisted approach* and the *deductive approach*.

First, the hermeneutic approach involves qualitative methods in which frames are identified using a detailed analysis of their links with broader cultural elements. These methods are often in-depth, well-documented and exceptionally detailed [50]. However, critics generally agree on the lack of quantification; In most studies, no or limited explanation is provided on how frames were extracted. As a consequence, these methods are criticised for being deficient in robustness, reliability and the danger of supporting the bias of researchers. "Researchers run the risk of finding frames they are consciously or unconsciously looking for" [50].

Secondly, the linguistic approach focuses on the selection, placement and structure of specific words and sentences to identify frames. These methods assume that frames in text are manifested in certain key words, stock phrases or sentences [20]. Frame elements are documented in a data matrix and therefore, these methods have the advantage of being systematic. However, this could also make the method rather complex in the case of large text samples. Additionally, the constitution of the elements to an actual frame can be unambiguous. [50].

Thirdly, the manual holistic approach conducts a qualitative pre-study to identify frames. Afterwards, these frames are coded as holistic variables in a *codebook*. A quantitative content analysis on all content using the codebook is used to extract all frames. Although these methods are more systematically than the hermeneutic approach, the same main problem arises: the methodology remains a black box because no or limited explanation is provided on how frames are identified. Additionally, manual holistic approaches risk that once frames are defined in the codebook, it might be difficult to observe the emergence of new frames [50].

Fourthly, computer-assistant methods are, as the name suggested, supported by a computer program. Within these quantitative methods, *dictionary-based* approaches are most common. In these approaches, previously defined words or combinations of words are related to content categories, or directly to frames [14]. The main advantage of these methods involves the possibility to analyse large amounts of content systematically. In an increasingly digital world, it is even arguable that these methods will become a necessity [14]. More advanced, *syntactical approaches*, are able to also capture the meaning of sentences. However, the need of these methods to previously define dictionaries, rules or categories can be both time-consuming and sensible to bias of the subjective conceptions limited domain knowledge of researchers [14]. To overcome these issues, Burscher et al. propose a *supervised machine learning* approach. Such a method is able to predict the frame present in an article, after it is being trained on a *training set* of labelled data. In the training set, articles are represented as a *bag of words*, a count of the occurrence of every word. The label of an article in the training set includes one of the four predefined generic frames, coded by 30 human coders. The classifier of Burscher et al. is able to outperform a baseline in which the frames are randomly distributed over the articles. The main advantage of these methods includes the increased efficiency and

accuracy compared to dictionary-based methods.

In contrast to the previous four inductive methods, the final category includes deductive content analysis of media frames. As mentioned before, these methods use predefined frames in their analysis, often derived from literature. However, this is also seen as the major limitations of the approach; deductive methods require a clear idea of the frames related to an certain topic before any analysis. But how to be sure the whole topic is covered and moreover, emerging frames will be recognised?

As a general conclusion after reviewing these five approaches in literature, Matthes and Kohring argue that the identification of frames often falls into a methodological black box. Thereby, the main issue includes the ambiguity of "which elements should be present in an article or news story to signify the existence of a frame" [50]. To overcome this problem, a new method is proposed based on framing elements in the the definition of Entman [20]. Thereby, it is assumed that a frame is a specific, systematic clustering of these frames element. During an experiment with around 1000 articles of The News York Times on biotechnology, Matthes and Kohring perform a *hierarchical cluster analysis* on the identified frame elements to extract the main frame of an article. As a consequence of the focus on frame elements instead of frames, the method could take advantage from improved reliability and less bias [50]. Likewise, Baden and Springer conceptualise viewpoint diversity by means of a diversity of frames, according to the four framing functions described by Entman [5, 20]. Thereby, frames are seen as concrete contextualisation for a specific issue, while *interpretative repertoires* are considered as generalised worldviews. Therefore, Baden and Springer go a step further by stating that frames are only diverse when they relate to different interpretative repertoires [5].

Overall, it can be concluded that current research on frames in news content is focused on the extraction and identification of frames, rather than the diversification based on framing metadata. However, the definition of Entman including four aspects of framing can be an interesting starting point [20]. The definition is most widely used among research on framing and provides concrete aspects of the concept, which would enable the translation to computational equivalents. Furthermore, the study by Matthes and Kohring in which frames are clustered using their framing aspects can form the basis for this study as well [50]. Among other scholars, the approach using framing aspects is seen as promising [88]. Different from their method, however, framing elements in this study will not be clustered but form the input to a diversification method. Although news content could include multiple frames according to Entman, this study will follow the research of Matthes and Kohring by scoping on framing aspects related to the main frame [50, 20].

2.4.4. FRAME BUILDING

Studies on the process of frame building question what and how certain factors influence how journalists and news organisations frame an issue [17]. Although a majority agrees on its importance, this part of the framing process has received the least attention in literature [17]. When addressing this part of the framing process, most scholars use media frames as dependent variable [88]. Scheufele identified at least five factors of influence: social norms and values, organisational pressures and constraints, pressures of interest groups, journalistic routine and ideological or political orientations of jour-

nalists [73]. Thereby, the focus is on the structural higher-level elements, rather than effects of the individual journalist [88]. Also, only recently, more attention has been paid to the variation of frames over time, across outlets and countries. For example, studies evaluate the difference of news frames on a given issue between different countries. Generally, it has been demonstrated that differences in the media system influence how a certain issue is framed [88]. Finally, some studies evaluated the influence of political actors. This is especially interesting regarding the major role frames play in the exertion of political power [20]. Since frames highlight some aspects of an issue while obscuring others, politicians have to compete with each other and journalists over news frames [20]. For example, Entman indicated how, during the Iraq war, media only echoed two closely related frames from the political elite. Following Entman's definition, the frames suggested only different remedies, wait or fight, for the issue. However, other, truly different frames, were not covered at all [20].

2.4.5. FRAME SETTING

Studies on the process of frame setting evaluate the effect of media frames on the audience. Within the whole research field, this part of framing is probably the most-widely studied [88]. This includes different domains, including sociology, psychology and communication. Sociology studies on framing can already be found in the 80s and address, for example, the role of frames in social movements [78]. However, most studies on frame setting use a *cognitive approach*, in which the effects of frames on the individual are analysed through quantitative experiments. The origins of this approach can be found in the social psychology, of which the work by Kahneman and Tversky form the best example. In their study, two participant groups were described a scenario of a virus outbreak among a population of 600 people. The first group was given the following two choices:

A 200 people will be saved.

B one-third probability that 600 people will be saved. two-third probability that no one will be saved.

The second group, however, were given these two choices:

A 400 people will die.

B one-third probability that no one will die. two-third probability that 600 people will die.

Among the first group of participants, 72% chose option A against 28% that chose option B. However, the second group showed opposite results: 22% chose option A and 78% option B [36]. Thereby, the relatively simple experiment indicates the significant influence of the framing of an issue on people's evaluation and choices [73].

Currently, framing has become the prominent media-effect theory in mass communication studies and many scholars evaluate how different frames influence a wide range of attitudinal and cognitive variables, such as opinions, political cynicism and emotions [88]. Thereby, the role of various aspects, such as personal characteristics, the framing

issue, or source, is considered. In particular the influence of individual political knowledge is often evaluated [88].

Although the focus of this study is not on framing effects, the overall conclusions of studies addressing this part of the process is relevant. In general, it can be seen that frames in news media have significant influence on their audience. For example, Porto have shown that when news coverage is restricted to a limited range of interpretive frames, more citizens interpret political events and issues according to the dominant frame [64, 65]. Also, Sniderman and Theriault showed that citizens tend to deviate farther from their core values when they receive uncontested single frames than when they receive balanced frames [77]. Therefore, it can be argued that these results support the need for the inclusion of a wide diversity of frames to ensure a diverse media landscape.

2.4.6. SUMMARY

The section can be summarised by the following points:

1. Research on framing can be described using the integrated process model of framing. Within this process, the manifestation of frames in the news is most interesting in the light of this study.

2. Within research on frame manifestations in news content, the definition of Entman is most commonly used [20]. This definition states that framing includes the selection of "some aspects of perceived reality and make the more salient in a communicating text, in such a way as to promote a particular definition of a problem, causal interpretation, moral evaluation and/or treatment recommendation for the item described" [20].

3. Current research on manifestations of frames in the news focuses on frame identification or extraction, rather than diversification based on framing metadata. However, the work of Matthes and Kohring, who evaluate frames using the framing aspects described in the definition of Entman, is seen as promising and can form an interesting starting point for this study as well. In contrast, the framing aspects will not be used for clustering but as input to a diversification method. Among other scholars, the approach using framing aspects is seen as promising [88].

4. In literature, the distinction between issue-specific and general frames is often made. However, this distinction will not be relevant for this work, since no frame identification will take place, only diversification based on framing aspects.

5. Although news content could include multiple frames according to Entman, this study will follow the research of Matthes and Kohring by scoping on framing aspects related to the main frame [50, 20].

2.5. NATURAL LANGUAGE PROCESSING TECHNIQUES AND TOOLS

This section provides a brief introduction into *natural language processing* (NLP). NLP methods aim to automatically process and understand human language and have a wide application in several domains, such as language translation, social media analysis and

customer sentiment monitoring. First, an overview of preprocessing techniques are provided. Afterwards, the most relevant NLP-methods for this work are described. Finally, multiple NLP-toolkits that have implemented these methods are implemented.

2.5.1. Preprocessing Techniques

Generally, *preprocessing* is the essential first step of every natural language processing pipeline. The task involves the transformation of the raw input format into an optimal form for a nlp-model. The most common techniques are described below:

Cleaning

The prepossessing task called *cleaning* involves several smaller sub-tasks to remove noise from the input format. A typical example includes the removal of HTML-tags when the input is scraped from web content. Besides, examples of cleaning tasks involve the removal of extra white space, the conversion of all task to lowercase characters, removal of special characters or the expansion of contractions [71].

Stemming and Lemmatisation

After the cleaning process, the preprocessing step of stemming or Lemmatisation is often applied. For grammatical reasons, different forms of a word are applied in content, such as walk, walks and walking. Also, different words can have the same meaning, such as democracy, democratic and democratization. Stemming and Lemmatisation both aim to transform any word to its base form. For example, "am", "are" and "is" will be transformed to "be".

An essential step in the process of stemming and lemmatisation includes *tokenisation*. Thereby, a document is split into words, such that it can be used by more advanced methods.

Stemming and lemmatisation differ, however, in their transformation methods. Stemming usually includes an heuristic process that chops the ends of words, thereby hoping to obtain the base form [74]. In contrast, Lemmatisation uses a vocabulary and morphological analysis of words to retrieve the base form of a word. In general, stemming is chosen for its speed, whereas Lemmatisation is preferable for quality reasons [74, 71].

Stop words

Finally, a regularly applied preprocessing step in natural language processing involves *stop word removal*. Stop words are considered to have little or no significant meaning but do appear very frequent in a text. Examples include, "we", "the" or "an". Stop words are removed using a dictionary lookup. Thereby, different dictionaries exist, such as the one in the NLTK python nlp library.

2.5.2. Relevant Natural Language Processing Techniques

Below, an overview of the most relevant natural language processing techniques is provided:

Part-of-Speech tagging

Parts of speech (POS) tagging includes the task of assigning lexical categories to words in a sentence, based on their syntactic context and role [71]. The most commonly present

tags include noun, verb, adjective and adverb, but more categories exists. Moreover, each tag can be divided into sub-tags. For example, a noun can be divided into singular nouns, singular proper nouns and plural nouns. Many other natural language processing techniques rely on POS-taggers, such as speech recognition, question answering and word sense disambiguation [37].

Different methods have been proposed for the task of part-of-speech tagging. Although early methods used a rule-based approach, current work can be mainly divided into supervised and unsupervised approaches. Supervised classifiers learn a mapping from human-defined features, while unsupervised approaches learn from the input representation themselves [37].

NAMED ENTITY RECOGNITION

Named Entity Recognition (NER) includes the task of identifying named entities, such as persons, locations or organisations [92]. Thereby, these methods provide information of significant relevance within a context. The first types of approaches were based on handcrafted rules, lexicons, orthographic features and ontology's. Afterwards, feature-engineering and machine learning models became increasingly popular. In the last decade, semi-supervised and unsupervised approaches have gained more momentum. Thereby, most models are specialised for a certain language. Generally, neural networks outperform feature-engineering based methods [92].

TOPIC MODELING AND LATENT DIRICHLET ALLOCATION (LDA)

Topic models are widely studied and applied in several domains, such as software engineering, political science and linguistic science [35]. These models aim to provide insights in the meaning of a text by discovering abstract topics within the document. Thereby, a document is assumed to be a mixture of topics, where a topic is a probability distribution over words. Concretely, topic models thus assume that if a topic is represented in a document, specific words related to the topic are included more frequently. Because of that, topic models are able to discover patterns of word-use across different documents.

Latent Dirichlet allocation (LDA) was introduced in 2013 by Blei, Ng, and Jordan and is currently one of the most popular topics models [35, 12]. LDA is an unsupervised generative probabilistic method. Each document is modeled as a mixture of topic, which is assumed to be a discrete probability distribution over a set of words. Thereby, the order of words is not taken into account; LDA uses a *bag-of-words* representation as input, in which every word is represented only by its frequency in the document. Compared to other topic models, such as *Latent Semantic Indexing* (LSI) and *Probabilistic Latent Semantic Indexing* (pLSI), LDA enables multiple topics per document and improves on the overfitting issue the previous methods were struggling with.

SENTIMENT ANALYSIS

Sentiment analysis involves the task of assigning a so-called *polarity* score to a text, based on people's opinions, attitudes and emotions towards and entity, including individuals, events or topics [52]. The analysis can be performed at different levels, including

the document, sentence and aspect level. The first two approaches aim to assign a polarity score that represents the sentiment of the document or sentence, while aspect-level sentiment analysis aims to retrieve a score with respect to a certain aspect or entity [52].

Generally, two types of approaches can be distinguished: machine learning approaches, including supervised and unsupervised methods, and *lexicon-based approaches*, including dictionary-based and corpus-based methods [52]. Thereby, machine learning approaches use linguistic features, while lexicon-based approaches rely on collections of recompiled sentiment terms.

SUGGESTION MINING

Suggestion mining involves the task of retrieving sentences that contain advice, tips, warnings and recommendations from the opinionated text [56]. The research domain is fairly young and has mainly been involved in the context of customer reviews. In this context, traditional methods mostly focused on the sentiment of customer towards a certain product of service. Suggestion mining was proposed to enhance these methods by extracting actionable feedback [87].

The majority of the proposed methods use rule-based approaches, in which linguistic patterns, such as keywords and POS-tags, are used as an indicator for suggestion sentences [87, 66, 53, 27]. Recently, machine learning approaches have become popular, including *support vector machines*, *Hidden Markov Chains* and *factorisation machines* [57]. Additionally, the first deep learning approaches have been proposed in this research domain [57].

To our best knowledge, only one study can be found that aims to evaluate suggestion mining approaches for other types of content [57]. For that purpose, general applicable rules of previous methods are aggregated. However, training machine learning models for general purposes is still very limited since most available datasets fully focus on customer reviews.

SEMANTIC ROLE LABELING

The task of semantic role labeling was firstly proposed in 2002 by Gildea and Jurafsky and involves the task of recover the predicate-argument structure of a sentence, to determine essentially "who did what to whom", "when" and "where" [30, 25]. The task has become important in many natural language processing application, since it is able to provide relational information between different aspects and entities in a sentence.

Since the Semantic Role Labeling is closely related to syntactic methods, such as part-of-speech tagging, traditional methods used information from syntactic models as a starting point. This, however, also restrains the models [80, 45]. Therefore, *neural sequence models*, such as *Long short-term memory* (LSTM), have become more popular recently. Generally, these methods have proven to outperform the traditional approaches.

KEYWORD EXTRACTION

Keyword extraction involves the task of identifying key terms, phrases, key segments or keywords from a document that can appropriately represent the subject of the document [9]. The technique has mainly been proposed as an answer to the increasing content available on the internet. Capturing the essence of a document in a small representation can benefit many applications, such as text summarisation, classification and clustering.

Two major approaches can be distinguished: *keyword assignment* and *keyword extraction*. In keyword assignment, an analysis of the content is used to assign keywords from a predefined taxonomy to a document. In contrast, keyword extraction retrieves keywords that are explicitly mentioned in the document [9].

2.5.3. NATURAL LANGUAGE PROCESSING TOOLS

Because of the wide application of natural language processing techniques, different toolboxes exist. These toolboxes provide a variety of methods using, for example, an API-service or programming library. For a couple of popular toolboxes, a short summary is provided below:

NLTK

NLTK is among the most popular toolkits that are around. The toolkit involves a set of Python modules that provide basic classes for data representation, interfaces for performing NLP tasks and implementations of these tasks. The toolkit provides an implementation for common nlp-methods, such as tokenisation (word and sentence), stop word removal, stemming, lemmatisation, POS-tagging and sentiment analysis.

STANFORD CORENLP

Stanford coreNLP is a nlp-toolbox developed and maintained by the Stanford NLP group [44]. The toolbox runs on Java but offers many different interfaces for other programming interfaces, such as Python. Also, it supports different languages including Arabic, Chinese, English, German, French and Spanish. Stanford coreNLP includes a variety of nlp-methods, such as tokenisation, POS-tagging, named entity recognition and sentiment analysis.

2.5.4. IBM WATSON NLP

IBM provides a natural language processing toolbox using a cloud-based API service. After sending a text document or web-page as input, IBM Watson returns semantic features, such as entities, keywords, semantic roles, sentiment. The system supports 13 different languages. However, not all features are supported on all languages.

2.5.5. LAMACHINE

LaMachine is a unified software distribution for natural language processing in Dutch. Thus, it is not a tool itself but provides a distribution of several other tools. The tool includes all nlp-tools by the Centre for Language and Speech Technology from the Radboud University in Nijmegen, such a tokenisation, POS-tagging and named entity extraction. Additionally, other methods from Dutch research groups and third party software are including.

It can be seen that these toolkits include a wide variety of commonly-used natural language processing techniques. Although the performance of the individual method will not match the most-recent state-of-the-art models, the methods are generally seen as well-performing and are used in several different research domains. According to Pinto, Gonçalo Oliveira, and Oliveira Alves, using these toolkits enables the development of more powerful application without having to start from scratch [62].

Since this study aims to enhance viewpoint diversity using framing aspects, multiple nlp-methods need to be combined to implemented the enrichment pipeline. Regarding the scope of the project, it has been chosen to put emphasis on the full pipeline rather than the optimisation of each individual method. Therefore, NLP-toolkits are assumed to be best suitable for this study.

2.5.6. SUMMARY

The section can be summarised by the following points:

- Natural language processing methods aim to automatically process and understand human language and have a wide application in several domains, such as language translation, social media analysis and customer sentiment monitoring.
- The task of preprocessing is an essential step in every natural language processing pipeline and involves the transformation of raw input data to a format that is optimal for a nlp-model. Examples include cleaning, stemming, lemmatisation and stop word removal.
- Relevant natural language processing techniques include part-of-speech tagging, named entity recognition, topic modeling, sentiment analysis, suggestion mining, semantic role labeling and keyword extraction. A short summary of these methods is provided above.
- Because of the wide-adaption of nlp-method in various domains, several toolboxes have been developed. These toolboxes include many popular nlp-methods and are freely accessible through programming libraries or API-services.
- Although the performance of the individual method will not match the most-recent state-of-the-art models, the methods are generally seen as well-performing and are used in several different research domains. According to Pinto, Gonçalo Oliveira, and Oliveira Alves, using these toolkits enables the development of more powerful application without having to start from scratch [62].
- Since this study aims to enhance viewpoint diversity using framing aspects, multiple nlp-methods need to be combined to implemented the enrichment pipeline. Regarding the scope of the project, it has been chosen to put emphasis on the full pipeline rather than the optimisation of each individual method. Therefore, NLP-toolkits are assumed to be best suitable for this study.

2.6. CONCLUSION

A literature study was performed to both gain insights in the research domain and evaluate related work. The findings related to the research questions are presented below.

RQ 1: *How is diversity defined in the context of news media? What conceptualisation can be used to diversify news recommendation?*

As described before, most studies on diversification in recommender systems define diversity as the opposite of similarity and propose methods that are based on topic diversity. These methods are thus, not directly applicable in the news domain. As a starting point for research on a novel diversification method for news media, a literature study was conducted on how diversity is defined in the context of news media and what conceptualisation can be used to to diversity news recommendations.

Diversity in news media is understood as as multiperspectivity or a diversity of viewpoints. Two main approaches to assess viewpoint diversity can distinguished: method based on content diversity and methods based on source diversity. Scholars generally agree that viewpoint diversity can only by achieved by fostering content diversity [89]. Among studies on content diversity, frames are generally seen as one of the most suitable conceptualisations of this concept. The definition of a frame by Entman is most commonly used [20]. This definition states that framing includes the selection of "some aspects of perceived reality and make the more salient in a communicating text, in such a way as to promote a particular definition of a problem, causal interpretation, moral evaluation and/or treatment recommendation for the item described" [20]. Current research on manifestations of frames in the news focuses on frame identification or extraction, rather than diversification based on framing metadata. However, the work of Matthes and Kohring, who evaluate frames using the framing aspects described in the definition of Entman, is seen as promising among other scholars [88]. Moreover, approaching frames by using framing aspects enables the translation of the concept in the social domain to its computational equivalents in the technical domain.

Therefore, frames and in particular the framing definition of Entman was chosen as suitable conceptualisation of diversity in the context of news media. Following Matthes and Kohring, the focus will be on the main frame of an article and no distinction between issue-specific and general frames will be made.

RQ 2: *What metadata can be related to this conceptualisation and which methods and tools can be used for the extraction of this data?*

Based on the answer to research question 1, metadata related to the four framing aspects as described by Entman needs to be extracted [20]. Therefore, Natural language processing Toolboxes will be used. These toolboxes include many popular nlp-methods and are freely accessible through programming libraries or API-services. Although the performance of the individual method will not match the most-recent state-of-the-art models, the methods are generally seen as well-performing and are used in several different research domains. Since this study aims to enhance viewpoint diversity using framing aspects, multiple nlp-methods need to be combined to implemented the enrichment pipeline. Regarding the scope of the project, it has been chosen to put emphasis on the full pipeline rather than the optimisation of each individual method. Therefore, NLP-toolkits are assumed to be best suitable for this study.

To understand which metadata could be related to each framing aspect, a focus group was organised with three experts in the field of journalism, communication and news media. Chapter 3 elaborates farther on the focus group.

RQ 3: *How can this metadata be combined to a measure for viewpoint diversity that can be used in a recommender system?*

The Maximal Marginal Relevance (MMR) algorithm is assumed to be most suitable for the diversification method. Thereby, a similar approach is taken as the most comparable work by Tintarev et al. [82]. However, in this study metadata related to the four framing aspects, as described by Entman, will be used [20].

Depending on which metadata can be related to each framing function, a distance function is needed to measures the diversity of two articles in terms of this framing aspect. Together, the distance functions of all four framing aspects form the diversity measure that can be used by the MMR-algorithm.

3

Focus Group

3.1. Introduction

As described in conclusion of the previous section, the conceptualisation of viewpoint diversity using the four framing aspects, including a problem definition, causal attributions, moral evaluation and treatment recommendations, described in the definition of Entman is assumed to be most suitable for this work [20]. Thereby, the second research question was partly answered:

RQ 2: *What metadata can be related to this conceptualisation and which methods and tools can be used for the extraction of this data?*

However, what metadata can be related to these four framing aspects is still unknown. The focus group is considered as a starting point for the translation of framing theory as understood in social sciences to computational equivalents that can be used in this study. The aim of this focus group is to gain insight in how the framing functions related the main frame of an article manifest in its content and how they can be recognised.

3.2. Setup

In this section the setup of the focus group is discussed. First, some information about the participants is given. Afterwards, the procedure that was used during the sessions is described.

3.2.1. Participants

For the focus group, three experts in the field of news article analysis and framing were invited. All experts have a background in journalism, communication or news media and have multiple years of relevant work experience. The participation was on voluntary basis and there was no incentive offered.

3.2.2. PROCEDURE

To be able the obtain in insights as described in the objectives, two tasks were set out to all experts. First, the they were asked to perform a framing analysis on a news article. Secondly, the results were discussed, together with some general questions on news article analysis and framing, during a review session.

FRAMING ANALYSIS

During the framing analysis, every expert was asked to analyse a news article on the topic of the farmer's protest in the Netherlands, based on the four framing functions as described by Entman [20]. In particular, the participants were given the task to highlight parts of the article, such as words clauses or sentences, that can be related to one of the four framing functions of the main frame of the article. Thereby, all four framing functions were assigned a specific color. Figure 3.1 provides an example of an highlighted section.

Figure 3.1: Example of the framing highlighting task during the focus group session

REVIEW SESSION

After the framing analysis, the results were discussed with the expert during a review session. For every highlighted part, the participant was asked why this part of the article could be related to one of the four framing functions. In addition, the results were used as input to a broader discussion on news article analysis and framing in news. Thereby, the following questions were discussed:

- Can you describe the procedure you followed while analysing the framing functions in the article?
- Within this procedure, are there any general methods, heuristics or tools you use when analysing an article?

- Regarding the individual framing functions, can you derive any generalities or patterns in the way they manifest in opinionated news articles?

3.3. RESULTS

The procedure as described in the previous section has been followed with all three experts. The following two main results were obtained:

3.3.1. MAIN HEURISTIC: ARTICLE STRUCTURE

During the review sessions, all experts indicated that the structure of the news article was used as main heuristic in finding the framing functions related to the main frame. Thereby, they pointed out that for the most common types of articles, including background analysis and opinion pieces, strong journalistic manners on how an article should be structured prevail. Therefore, this heuristic further analysed with the experts according to the four framing functions:

1. **Problem Definition**
   In background analysis and opinion pieces, the first part of the article is often used to present the main problem that the author is addressing. the first part can include title, lead and the first x paragraphs. This is supported by work on manual frame analysis by Kroon et al. [42]. The number of introductory paragraphs, represented by the variable x, can be different per source, author or article.

2. **Causal Attributions + Moral Evaluation**
   The body of an article, all paragraphs except introductory and concluding paragraphs, are then used to analyse the main problem. Thereby, different factors that contribute to the problem under investigation are described, often accompanied with an evaluation of these factors. This can be easily matched with the definition of two framing functions: causal attribution of a frame relates to the forces creating the problem, while the moral judgements evaluate the causal attribution and their effect [20].

3. **Treatment Recommendations**
   Treatment recommendations can be seen as suggestions to improve on or solve the issue as described by the problem definition of the main frame. If an author provides any suggestions to the problem of the main frame, they can normally be found in the concluding paragraphs, according to the focus group members. Thereby, the author is referring back to the main problem, which is seen as orderly way to conclude a story.

A visual summary of these points can be found in Figure 3.2. Note, however, that this structure is not strictly applicable to every article. It can only be seen as an heuristic. Moreover, it only applies for background analysis and opinion pieces. Other types, such as interviews, are structured differently.

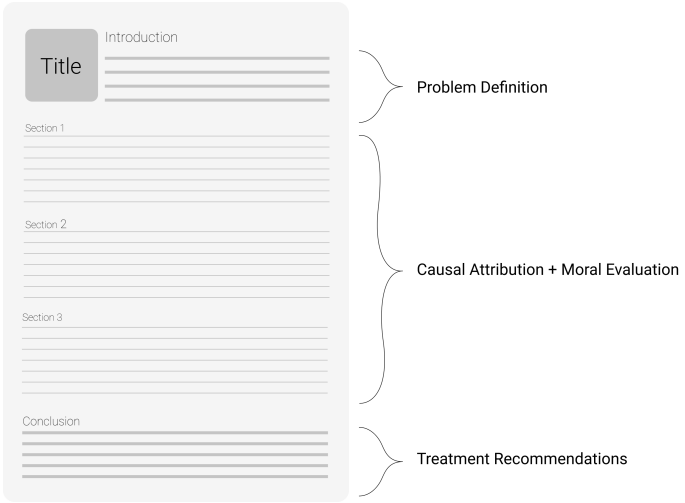


Figure 3.2: Overview of most important insight of focus group

3.3.2. MANIFESTATION OF FRAMING ASPECTS

Additionally, the frame analysis task and review sessions provided two insights into the manifestation of the four framing functions in text:

- **Level of analysis**
  The results of the highlighting task indicate that individual framing functions related to the main frame of an article can normally be found within one paragraph. In there, the manifestation of the framing functions can be different, including short word combinations, clauses, sentences and a combination of sentences. Additionally, a paragraph can include multiple framing functions, but words, clauses and sentences generally represent a single framing function.

- **Context**
  Finally, the importance of context during frame analysis became clear. Although the manifestation of individual framing functions related to the main frame can normally be found at the level of a single paragraph, context outside the paragraph can be needed to reveal a framing function. This includes context within the article, on the level other paragraphs or the total article, but also context outside the article, such as general knowledge.

3.4. CONCLUSION

As described in the introduction of this chapter, the focus group was organised to gain insight in how the framing functions related the main frame of an article manifest in its content and how they can be recognised. Thereby, the following research question was addressed:

RQ 2: *What metadata can be related to this conceptualisation and which methods and tools can be used for the extraction of this data?*

Based on the results of the focus group, the following conclusions related to this research question can be drawn. First, the common structure of background analysis and opinion pieces can be used as main heuristic for finding the four framing functions described by Entman [20]. Thereby, the problem definition framing function related to the main frame can, according to the heuristic, be found in the title, lead and first x paragraphs of an article. This variable differs per source, author or article and thus, needs to be optimise during an offline evaluation. Additionally, the causal attribution and moral evaluations framing functions related the article's main frame are worked out in the body of the article, which includes all but introductory and concluding paragraphs. Finally, if an author includes treatment recommendations for the main problem under investigation, this is regularly worked out in the y concluding paragraphs. Similar to the number of introductory paragraphs, this variable can variate and thus, needs to be optimised during an offline evaluation.

Additionally, the paragraph-level is chosen as most suitable level of analysis. In there, framing aspects can manifest in short word combinations, clauses, sentences and a combination of sentences. Also, context can be important for framing analysis. During to time limitations, however, it was decided to consider context related aspects as out of scope.

4

METHODOLOGY

4.1. INTRODUCTION

Based on the conclusion of the previous chapter, the common structure of the most common types of articles can be used as main heuristic in finding the four framing functions described by Entman [20]. This chapter discusses the final choice for the metadata extraction, based on this heuristic. Therefore, different setups, implemented using NLP-toolkits, were considered. Additionally, the final choice and considerations of a global viewpoint diversity measure based on this metadata is described. Finally, the MMR reranking algorithm based on the viewpoint diversity measure is described. The chapter starts by an overview of the total pipeline.

4.2. OVERVIEW OF TOTAL PIPELINE

Figure 4.1 provides an overview of total diversification pipeline:

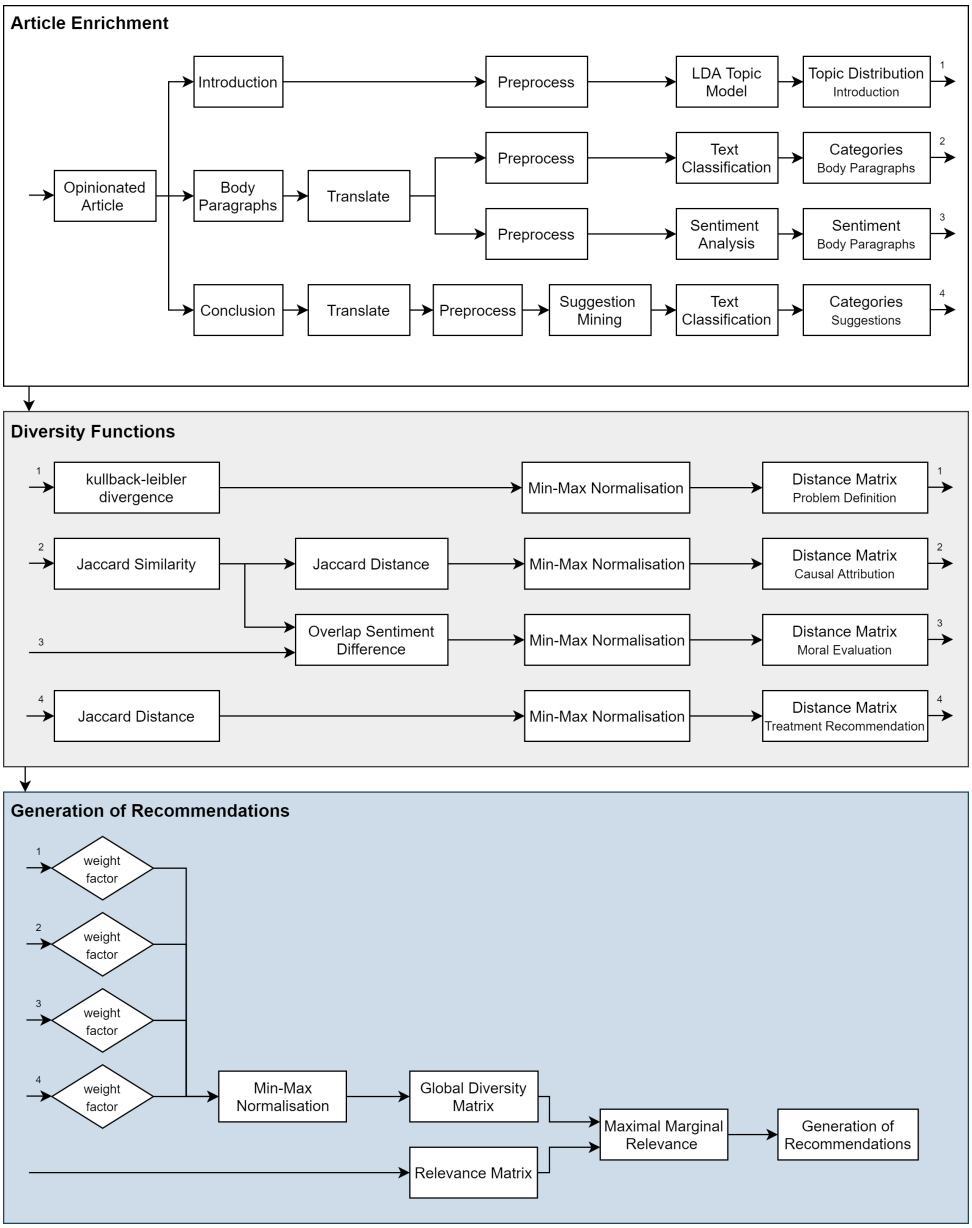


Figure 4.1: Overview of total pipeline

Given an article on a certain topic and a dataset of articles on the same topic, the pipeline is able to rerank a list of article recommendations from the dataset that are both relevant and viewpoint diverse compared to the input article. The process can be described by the following three sequential steps:

1. **Article Enrichment**
   First, every article in the set including the article for which the recommendation list will be generated will be enriched. Hereby, metadata related to each framing function as described by Entman will be extracted. A detailed description of this step can be found in section 4.3.

2. **Diversity Functions**
   Secondly, the information derived from the extraction process is used to calculate a diversity measure for each nonequal article combination in the set. The final diversity measure of two articles is composed by the weighted sum of the individual distance functions related to each framing functions. A detailed description of this step can be found in section 4.4.

3. **Reranking of Recommendation list**
   Finally, the distance measure in combination with a relevance score for every article combination is used to rerank the set of recommendations for a given article. Thereby, the Maximal Marginal Relevance (MMR) algorithm is used. A detailed description of this step can be found in section 4.5.

4.3. EXTRACTION OF INDIVIDUAL FRAMING FUNCTIONS

The first of the pipeline includes the extraction of information related to the main frame of every article. As described in the previous section, this is approached by individually extracting the metadata related to each of the four framing functions. This section describes the pipeline related to this process for each function. Section 4.3.1 describes the extraction pipeline for metadata related to the problem definition of the main frame. Section 4.3.2 elaborates on the combined pipeline for the causal attribution and moral evaluation. Lastly, section 4.3.3 explains the process related to the treatment recommendations of the main frame.

4.3.1. PROBLEM DEFINITION

The extraction of metadata related to the problem definition function of the main frame of an article is illustrated in Figure 4.2. The figure provides an overview of every step in the enrichment process. Each block contains both the type of operation that is being performed and the related tool or model. The justification of the choices related to these steps can be found in section 4.3.1 and 4.3.1. Within the enrichment process, three phases can be distinguished:

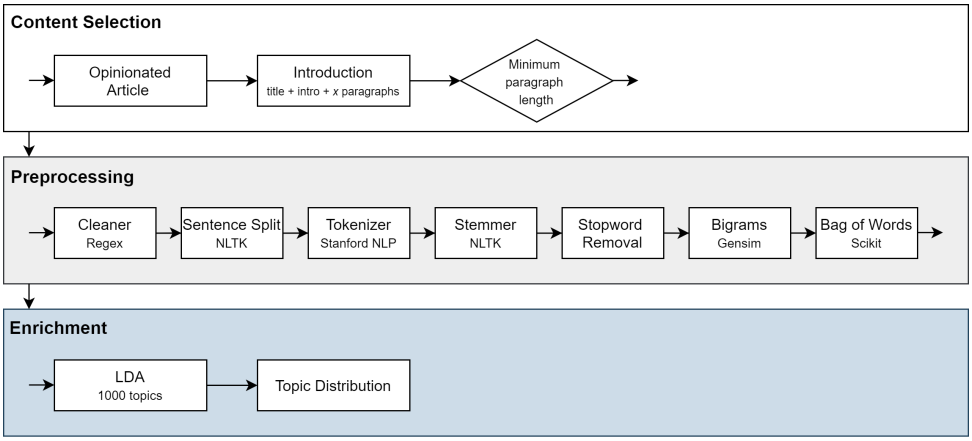


Figure 4.2: Overview of pipeline to extract metadata related to the problem definition of the main frame

1. **Content Selection**
   At the start of the pipeline, the content of interest is selected from the full article text. Based on the conclusions from the focus group described in chapter 3, the title, possible lead and first x paragraphs are extracted for this framing function. As earlier mentioned, this variable will be optimised during the offline evaluation. As a final step, the body paragraphs are filtered to have a minimum word length.

2. **Preprocessing**
   The selected introductory content will be preprocessed, such that it can be handled by the LDA-model. First, a cleaning process removes the most common oddities. This includes the elimination of HTML-tags and the removal or replacement

of erroneous punctuation. Secondly, the cleaned text will be split into separate sentences by the NLTK sentence splitter. These sentences form the input for the NTLK tokenizer, which splits the sentences into lowercase tokens, such as words or punctuation. Fourthly, the NLTK stemmer removes the morphological affixes from words, resulting in the root or base form of every word. Afterwards, the lemmas will be filtered, such that all stopwords, punctuation and short tokens (less then 3 characters) will be removed. Thereafter, the Gensim-phraser will search and combine any common double-word expressions. Finally, the processed tokens will be transformed into a bag-of-words representation. This will be the input for the LDA-model.

3. **Enrichment**
   Lastly, the preprocessed data is delivered to the LDA-model. This model is Blendle's standard LDA-model and is used to produce the daily article recommendations to all users. The output of the model includes a probability distribution over 1000 topics. The model is trained on 900k unique Dutch articles within a period of two years. Thereby, a token is included to the training set if it appears in more than 100 documents but in less than 80% of all documents. Additionally, only the 50k most frequent tokens are included. The model is thus trained on a sparse matrix of 50k by 900k.

JUSTIFICATION OF CONTENT ANALYTICAL VARIABLES

Metadata related to the problem definition of the main frame will be extracted using a topic model on the introductory parts of an article. Thereby, the method follows largely the work of Matthes and Kohring, who conceptualise the problem definition as a combination of the central issue or topic under investigation and the most important actor [50]. In their work, the central issue includes subtopics on biotechnology, such as biomedicine, cloning or regulations. Also, the main actor includes a more high-level conceptualisation such as politics or business, and should therefore not be confused with, for example, important entities in the text. Matthes and Kohring perform a pre-study to define the list of central issues and important actors [50]. In this work, it is assumed that because of the high-level conceptualisation of the problem definition, a topic model will be able to capture a significant part of the same information. Thereby, a topic model will allow a more general approach, without the need for any pre-study. As earlier mentioned, the content selection step is based on the main conclusion of the focus group session, which revealed the common structure of opinionated news article as most useful heuristic in framing analysis.

IMPLEMENTATION CHOICES

Concerning the topic-model, the choice has been made to use the latest version of the Blendle LDA-model. As mentioned before, the model is trained on a significant number (900k) of Dutch news articles from the same sources as those that will be used in this work. Therefore, it is assumed that this model will yield good performance for the purpose of our work as well. Since this model is trained on data that is preprocessed in a specific way, the preprocessing steps of this pipeline are chosen to match these steps exactly.

4.3.2. CAUSAL ATTRIBUTIONS AND MORAL EVALUATION

The extraction of metadata related to the causal attribution and moral evaluation of the main frame of an opinionated article is illustrated in Figure 4.2. The figure provides an overview of every step of the enrichment process. Each block contains both the type of operation that is being performed and the related tool or model. The justification of the choices related to these steps can be found in section 4.3.2 and 4.3.2. Within the enrichment process, three phases can be distinguished:

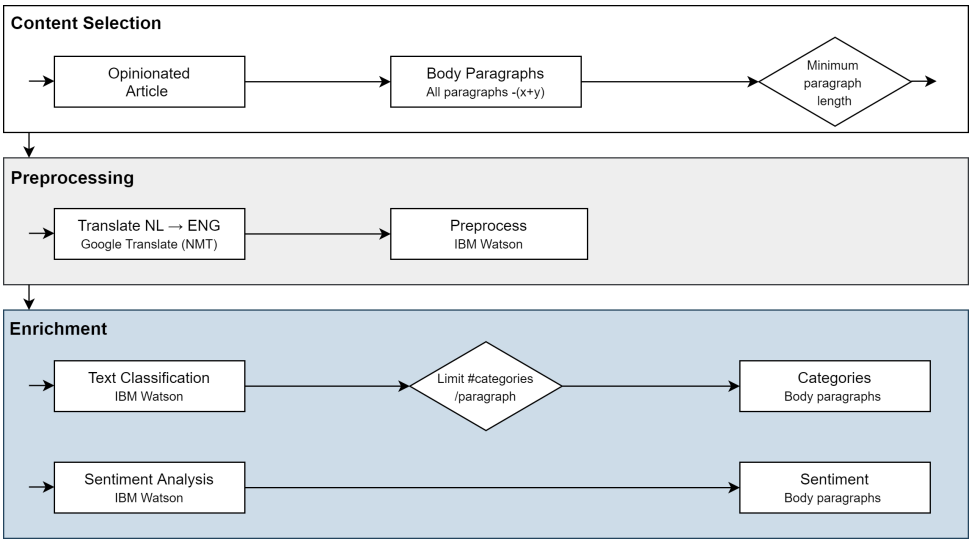


Figure 4.3: Overview of pipeline to extract metadata related to the causal attributions and moral evaluation of the main frame

1. **Content Selection**
   Similar to metadata extraction of other framing functions, the pipeline related to the causal attributions and moral evaluation framing functions of the main frame is initialised by the selection of the content of interest. As described in chapter 3, the content of interest for this task includes all paragraphs except the x introductory and y concluding paragraphs. These variables will be optimised during the offline evaluation. As a final step, the body paragraphs are filtered to have a minimum word length.

2. **Preprocessing**
   The preprocessing step is initiated by the translation of the selected body paragraphs from Dutch to English. This step is performed using the Google Translate API, including the NMT translation model. Secondly, the response from the API will be forwarded to the IBM Watson API to be preprocessed and enriched.

3. **Enrichment**
   After the body paragraphs are preprocessed by IBM Watson, two enrichment models are applied. First, every paragraph will be classified according to a pre-defined

five-level taxonomy based on the content of the paragraph. An example of this taxonomy for four levels can be found in table 4.1. For every category, an relevance score will be calculated as well. Using a limit on the number of categories per paragraph, the categories of all paragraph including their relevance score are combined to a document-level representation. The second model includes a sentiment analysis on paragraph level.

Level 1	Level 2	Level 3	Level 4
technology and computing	hardware	computer	portable computer
law, govt and politics	government		
science	social science	sociology	

Table 4.1: Example of five-level taxonomy of IBM Watson categories

JUSTIFICATION OF CONTENT ANALYTICAL VARIABLES

Following the definition of Entman, the causal attribution of a frame relates to the forces creating the problem, while the moral judgements evaluate the causal attribution and their effect [20]. According to Baden and Springer, attributed causes and evaluative judgement can be identified by answering the following questions: "what brought this focal concern about, according to the text? And how should that be evaluated?". From the discussion of the focus group session, described in chapter 3, it can be concluded that the body of an opinionated article is normally used by the author to elaborate on these aspects. Often, the author discusses per paragraph different actors that contribute to the main issue under consideration. This if often accompanied by a judgement of actor, based on the effect of their contributions to the issue. Based on these conclusions, it is assumed that framing metadata related to the causal attributions and moral evaluations of the main frame can normally be found side-by-side in its body paragraphs of an article.

To determine which content analytical variables can be used, two different options of the content analytical variables have been explored:

A **Entity and keywords extraction + entity- and keyword-level sentiment analysis**
In this option, relevant entities and keywords are extracted from the content of the body paragraphs. Additionally, the sentiment related to those terms will be extracted. The overlap of entities and keywords can be seen as a measure for the similarity of the causal attributions of both articles. Likewise, the difference in sentiment between the overlapping entities or keywords could be an indication for diversity of the moral attribution.

B **Paragraph text classification + paragraph sentiment analysis**
Instead of extraction detailed low-level information such as concrete entities, this option aims to relate the content of each paragraph to high-level categories using a text classification model. For example: instead of comparing articles on the presence of an entity such as "Mark Rutte", the paragraphs are compared using higher-level categories such as "political leaders". Additionally, the sentiment

related to those categories is measured using a sentiment analysis on paragraph level. Thereby, it is assumed that all categories in one paragraph can be related to the same sentiment score.

Both options for the content analytical variables have been tested on a small dataset of 21 articles on the Dutch nitrogen crisis. Although option A could enable a more detailed comparison between articles, the results indicated a very small overlap of entities and keywords between articles. As a result, no comparison in terms of the causal attribution and moral evaluation framing functions would be possible for most article combinations. In contrast, the comparison using high-level categories was possible for every combination of articles. Therefore, option B was chosen as preferable method.

IMPLEMENTATION CHOICES

The core of the pipeline is implemented using the IBM Watson natural language processing API. This choice is primarily based on the large number of features that is available in this toolkit. Because of that, the performance of different methods, including entity and keywords extraction, entity- and keyword-level and text classification, could be compared. Since this rich feature set is only available in English, the decision has been made to switch to English for extraction of information related to these framing functions. Thereby, it is assumed that the possible loss of information related to the translation step does not outweigh the benefit of using a rich toolkit. Also, it must be noted that the focus of this thesis is rather the implementation a well-performing full diversification pipeline, than the optimisation of each extraction task individually.

4.3.3. TREATMENT RECOMMENDATIONS

The extraction of metadata related to the causal attribution and moral evaluation of the main frame of an opinionated article is illustrated in Figure 4.4. The figure provides an overview of every step of the enrichment process. Each block contains both the type of operation that is being performed and the related tool or model. The justification of the choices related to these steps can be found in section 4.3.3 and 4.3.3. Within the enrichment process, three phases can be distinguished:

1. **Content Selection**
   Similar to the first three framing functions, the content of interested is selected first. For the metadata related to the treatment recommendation, this includes the y concluding paragraphs of an article. This variable will be optimised during the offline evaluation. As a final step, the body paragraphs are filtered to have a minimum word length.

2. **Preprocessing**
   Like the pipeline of the metadata related to the causal attribution and moral evaluation, the first of the preprocessing task includes the translation of the concluding paragraphs from Dutch to English. Again this is done using the Google Translation API, including the NMT model. Afterwards, the paragraphs are split into sentences using NLTK. These separate sentences form the direct input for an part of the enrichment process, indicated in figure by the number 2. Fort the other part of the

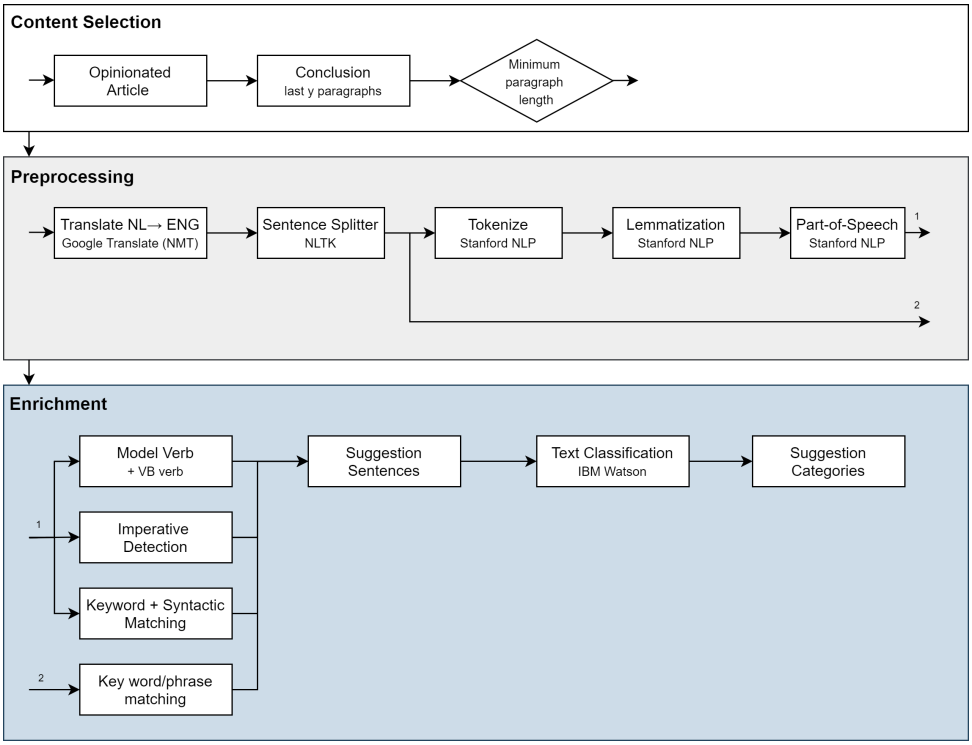


Figure 4.4: Overview of pipeline to extract metadata related to the treatment recommendations of the main frame

extraction process, some extra preprocessing tasks have to be performed. These include tokenisation, lemmatisation and part-of-speech tagging. All these tasks are implemented using Stanford coreNLP.

3. **Enrichment**
   The first column of the enrichment process includes all tasks related to the rule-based suggestion mining. The information from the tokenisation, lemmatisation and part-of-speech tagging is used in three of four rules. For the first rule, the sequence of a modal verb followed by the base form of an other verb is detected. Examples of this construction include "must stop" or "should prevent". An overview of all modal verbs that are considered can be found in table 4.2. The second rule aims to find imperative constructions, such as "stop spending money on the EU", by evaluating the presence of any verb in front of any word that could be the subject. Thereby, interrogative sentences are ignored. Lastly, the fourth rule uses the separate sentences from the preprocessing directly and performs a matching with a list of keywords and phrases.

   The output of all four rules is the same: the sentences of the concluding paragraph of each article that contain a suggestion. Finally, these sentences are classified by

IBM Watson in the same five-level taxonomy as the body paragraphs related to the causal attribution and moral evaluation. This step enables the comparison of suggestion sentences between articles.

JUSTIFICATION OF CONTENT ANALYTICAL VARIABLES

Following the definition of Entman, a treatment recommendation suggests remedies for the problems and predict their likely effect [20]. According to Baden and Springer, treatment recommendations can be found by answering the following question: "what is presented as a suitable course of action upon the issue?" [5]. According to the results of the focus group session, the conclusion of an opinionated article, regularly the last few paragraphs of an article, is the usual place for any suggestion by the author. Note, however, that it is not an obligation for an author to provide any suggestion; One can simply be critical about a subject without suggestion any solutions. If an author provides any, however, these suggestions are often provided in the concluding paragraphs of an article according to our focus group members.

As described in section 2.5, the research domain of suggestion mining involves the task of retrieving sentences that contain advice, tips, warnings and recommendations from the opinionated text [56]. Following this definition, it can be argued that the task of retrieving metadata related to the treatment recommendations of a frame largely overlaps with task described by this research domain. Therefore, proposed methods in this domain could potentially be interested for this work as well. As described in section 2.5, methods in this research field can be categorised as rule-based approaches, machine-learning approaches and deep-learning approaches [57]. Although machine-learning and deep-learning approaches yield the best performances, these methods are not directly applicable in domain of this study since these models are trained on domain specific-content, mainly product or service reviews [57]. Moreover, no datasets on suggestions in news articles are available [57]. Although some rule-based methods are focused on domain specific content, general solutions have been proposed. Therefore, rule-based suggestion mining is considered as the most suitable model for the extraction of treatment recommendations.

Concretely, two studies in the domain of suggestion mining choose a general rule-based approach [57, 56]. In the studies, the rules of domain-specific research that can be generally applied are aggregated. However, these rules have not been evaluated on the content of news articles. Therefore, there was a need to evaluate the performances of these rules on this type of content. For that purpose, a crowdsourcing platform was developed to create a ground truth dataset on suggestions in Dutch news articles. The experimental setup and results are described in section 6.2 and 6.3, respectively. From the results, the following set of rules is defined:

The suggestion-mining models is able to extract the sentences in the concluding paragraphs that contain information that can be linked to the treatment recommendation of the main frame. However, to compare this information between sentences of different articles an additional step is necessary. The following options were considered:

A **Text Classification**
In this option, the sentences are classified according to a predefined category taxonomy, such as "political leaders" or "Business plans". This choice would provide

Rule	Related work	Patterns
Modal verbs	[66]	Modal verb (could, must, may, shall, should, ought to) + Base form of verb (VB)
Imperative detection	[57]	Verb in front of noun
Keywords + Phrases	[66] [87] [53] [27]	suggest, recommend, hopefully, go for, request, it would be nice, adding, should come with, should be able, could come with, I need, we need, needs to, need to, would like to, would love to, I wish, I hope, hopefully, if only, would be better if, would that, I can has, do want, would like if, can't believe didn't, don't believe didn't

Table 4.2: Overview of rules that are included in the pipeline

an high-level insight on the differences between treatment recommendations related to the problem definition of the main frame.

B **LDA Topic Model**
In this option, the sentence is used as input (after the necessary prepossessing) to an LDA model. Similar to the previous option, the topic distribution of a LDA topic model could provide an high-level insight into how two treatment recommendation are be different.

C **Entity + Keyword overlap**
A more detailed comparison could be achieved by comparing the entities and keywords that are extracted from the suggestion sentences. A comparison between these terms could be implemented using a cosine similarity of the word vectors of the entities and keywords.

D **Role Extraction + one of previous options**
Finally, all previous options could be enhanced by a role extraction model. In that case, object and subject could be handled separately, which enables more detailed comparison. For example, this setup would in theory be able to distinguish between two different actors for which the same action is suggested. After the object and subject have been extracted, all three previously described options could be applied to obtain metadata that can be compared across articles.

All options for the content analytical variables have been tested on a small dataset of 21 articles on the Dutch nitrogen crisis. Thereby, the text classification model without role extraction appears to be the preferable choice. Like the experiment with entities and keywords in section 4.3.2, the overlap of these terms between articles was rather small. Therefore, only a few article combinations could be compared. Although the topic model was able to provide high-level information on the differences of suggestions, the information from the text classification has found to provide more valuable information. The

predefined categorisation on how topics relate to each other, enables a richer comparison between actors and topics in different treatment recommendations.

Finally, the quality of the output of the role extraction model appeared to be too low for the other models to perform well. Thereby, it must be noted that it was only possible for sentences containing modal verb construction to retrieve the exact subject and object related to the suggestion. For the other rules, too less information was available to understand which object and subject relate to the suggestion in the sentence.

IMPLEMENTATION CHOICES

Similar to the pipeline related to the causal attribution and moral evaluation, the IBM Watson natural language processing API is chosen as the main toolkit for this pipeline. Again, this choice is primarily based on large set of features available in the toolkit. For this pipeline, this includes entity and keyword extraction, role extraction and text classification. Moreover, the translation from Dutch to English was a inevitable step, due to the absence of any rule-based suggestion mining approaches in Dutch.

4.4. DISTANCE FUNCTIONS

After the metadata related to each framing function has been extracted, this information needs to be compared across all articles in the dataset. For that purpose, a distance function is needed for each framing function, which provides a measure for the distance between two articles in terms of the framing function. This section describes the pipeline related to this process for each function. The pipeline that implements the distance function for the problem definition, causal attribution + moral evaluation and treatment recommendations is described in section 4.4.1, 4.4.2 and 4.4.3, respectively.

4.4.1. PROBLEM DEFINITION

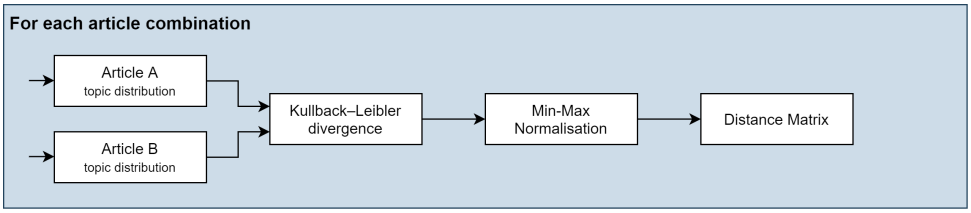


Figure 4.5: Overview of pipeline to calculate distance between two articles based on the problem definition metadata of each article

The pipeline that calculates the distance based on the problem definition metadata related to the main frame between two articles is illustrated in Figure 4.5. From the enrichment process described in section 4.3.1, the distribution over 1000 topics, computed by the Blendle LDA-model, are obtained for each article.

To compare pairs of articles in terms of this framing function, a distance function is required that is able to calculate the dissimilarity of two topic distributions outputted by the LDA-model. Since the LDA topic distribution involves a probability distribution, a statistical distance measure is needed. In this pipeline the Kullback-Leibler divergence

is implemented to calculate the distance between two topic distributions. The Kullback-Leibler divergence is one of the most commonly used statistical distance measures for LDA-models. Additionally, it is used in the comparable work of Tintarev et al. on viewpoint diversification [82]. Given two (discrete) topic probability distribution P and Q over a probability space of T of 1000 topics, the Kullback-Leibler divergence (KL) of P from Q can be calculated as follows:

$$KL(P||Q) = \sum_{t \in T} P(t) \log(\frac{P(t}{Q(t)}) \tag{4.1}$$

The idea behind function can be understood as follows: when the probability of a topic t in P is large, but the probability of the same topic in Q is small, the divergence is high. Vice versa, if the probability for a topic t is small in P but large in Q, the divergence is also large but smaller than in the first case [13]. Consequently, it is important to note that the Kl-divergence is not symmetrical, thus:

$$KL(P||Q) \neq KL(Q||P) \tag{4.2}$$

Therefore, both $KL(P||Q)$ and $KL(Q||P)$ are calculated for each combination of probability distribution P and Q. The divergence values are scored in a diversity matrix. Since the matrix holds two values for each combination of articles, except the the combination of the same article, the matrix is hollow. Thus, all diagonal entries are zero. Finally, the matrix is normalised using a min-max normalisation. The full process is illustrated in Figure 4.5

4.4.2. CAUSAL ATTRIBUTION AND MORAL EVALUATION

Similar to the metadata extraction, the distance functions related to the causal attributing and moral evaluation are partly overlapping. Therefore, the calculation is combined in a single pipeline, illustrated in Figure 4.6. Section 4.4.2 describes the process related to distance function of the causal attribution. Section 4.4.2 explains the process for the moral evaluation.

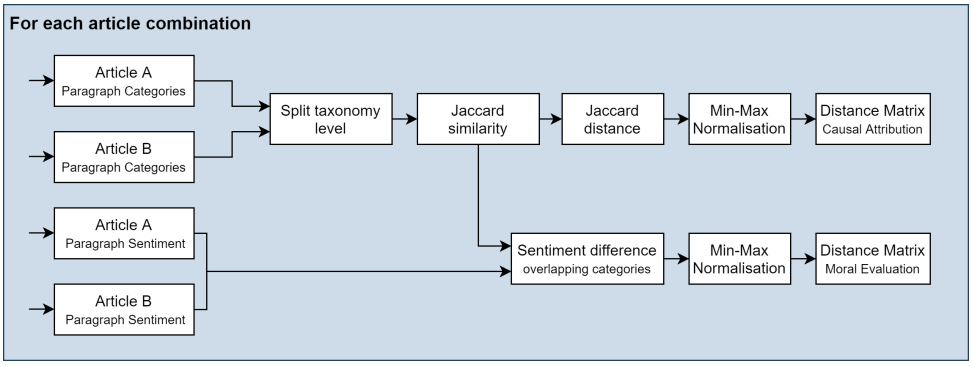


Figure 4.6: Overview of pipeline to calculate distance between two articles based on the causal attribution and moral evaluation metadata of each article

CAUSAL ATTRIBUTION

The five-level taxonomy categories extracted from the pipeline described in the previous section, must be compared to obtain a distance measure related to the causal attribution framing function of the main frame. To do so, the weighted Jaccard index was used, which includes a measure for the similarity (or diversity) of two sets [34]. The index will be calculated separately for each level of detail in the five-level taxonomy, such that weight factors per taxonomy level could be applied. Thereby, overlap in higher levels of detail can have a larger contribution to the overall similarity score. In the offline evaluation different weight factors per taxonomy-levels are compared. The total distance function can be described by the following steps:

1. **Split Taxonomy Levels**
   First, the categories of both articles are split up per taxonomy level, such that every article is represented by five distinct sets of categories. For example, the "business and industrial / business news" category containing two categories is divided in "business and industrial" and "business news". As described before, the separation enables different weights per taxonomy level, such that overlap in higher level categories can have a larger contribution to the overall similarity.

2. **Jaccard Similarity**
   Secondly, the Jaccard similarity is calculated for each taxonomy level. Thereby, the weighted version is used, such that the relevance index, returned for every category by IBM, is taken into account. To calculate the index, an input vector that represents the categories of each article is needed. This vector includes an entry for each category in $A \cup B$. If a category is represented in A, the entry contains the relevance score for that category in A. Otherwise, the value is zero. If a category is represented multiple times, the relevance scores are summed. From the vectors $A = (a_1, a_2, ..., a_n)$ and $B = (b_1, b_2, ..., b_n)$, the weighted Jaccard index for taxonomy level 1 can be calculated as follows:

$$J_1(A,B) = \frac{\sum_i min(a_i, b_i)}{\sum_i max(a_i, b_i)} \tag{4.3}$$

   Afterwards, the Jaccard of each taxonomy index can be combined using the corresponding weight factors:

$$J(A,B) = \sum_{j \in [1,2,...,5]} w_j \times J_j \tag{4.4}$$

   Note that the weight factors sum up to 1.

3. **Jaccard Distance**
   Thirdly, the Jaccard similarity derived in the previous step is used to calculated the Jaccard distance. This distance is defined as follows:

$$D_j = 1 - J \tag{4.5}$$

The distance is measured for each article relative to every other article in the dataset. The score is stored in a diversity matrix. Again this is an hollow matrix with zero values in the diagonal. Finally, the matrix is normalised using a min-max normalisation.

MORAL EVALUATION

As earlier mentioned, the distance function of the moral evaluation metadata is largely overlapping with the causal attribution distance function. From that pipeline, a similarity score between each paragraph combination of the articles is derived. This similarity measure is combined with the absolute difference of the sentiment score of the paragraph combination, such that highly similar paragraphs with diverse sentiment score will lead to high levels of diversity in terms of the treatment recommendation. The comparison between every paragraph $i \in A$ and $j \in B$ is made and summed up to a total diversity score. This can be described by the following equation:

$$D(A,B) = \sum_{i \in A} \sum_{j \in B} abs(S_i - S_j) \times J(A,B) \tag{4.6}$$

Finally, the diversity measures are normalised using a min-max normalisation and stored in a matrix.

4.4.3. TREATMENT RECOMMENDATIONS

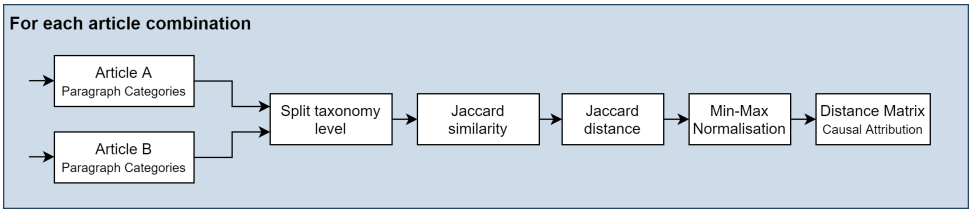


Figure 4.7: Overview of the pipeline to calculate distance between two articles based on the treatment recommendation metadata of each article

As described in section 4.3.3, the sentences derived from the rule-based suggestion mining will be classified by the IBM Watson text classification model. Therefore, the output of the enrichment process will include categories from the same five-level taxonomy as the output from the enrichment process of the causal attribution metadata. Therefore, a similar distance function is used for the metadata related to the treatment recommendation: for each taxonomy level, the Jaccard distance is calculated. These values are combined to one distance measure using the corresponding weight factors. Details of this calculation can be found in section 4.4.2, which matches the distance function exactly.

4.5. RERANKING OF RECOMMENDATION LIST

Below, the final step of the pipeline, which reranks the recommendation list, is illustrated. Within this pipeline, three steps can be distinguished: the calculation of the global diversity measure is described in section 4.5.1, the calculation of a relevance score

between every pair of articles is described in section 4.5.2 and the reranking of the list is explained in section 4.5.3.

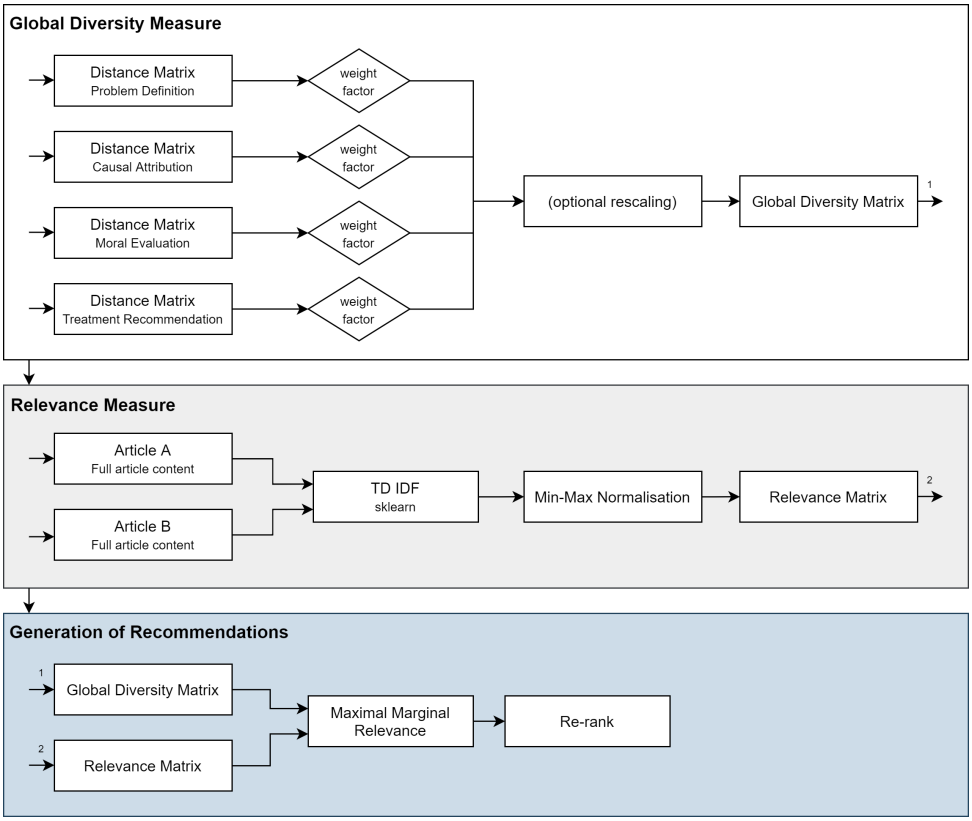


Figure 4.8: Overview of the pipeline to re-rank article recommendations

4.5.1. GLOBAL DIVERSITY MEASURE

First, the four distance functions described in the previous section are combined to a global distance measure for each unequal article combination in the dataset. This value is simply calculated as the weighted sum of the distance functions. Thus, the diversity measure $Div(i, j)$ between two articles i and j is sum over the distance functions $d_k(i, j)$ times the corresponding weight factors w_k:

$$Div(i||j) = \sum_{k=1}^{n} w_k d_k(i||j) \tag{4.7}$$

Afterwards, an optional re-scaling must be applied. This is a necessary step for an article combination if the distance function can not be calculated for at least on of the framing functions. For example, this happens regularly for treatment recommendation

framing function since many articles do not provide any suggestions to the main problem under investigation. According to Entman, a frame does not have to comprise all four framing functions [20]. Based on that, it assumed that the absence of any framing function in the comparison of two articles should not automatically imply a smaller diversity score. Therefore, the average value of the framing function over all article combinations will be used, if for a combination of articles no value can be calculated.

4.5.2. RELEVANCE MEASURE

Secondly, a relevance score is calculated for each article combination. Since this work focuses on the implementation of measure for viewpoint diversity rather than a relevance measure, the choice was made to implement this score using the a simple frequency–inverse document frequency (tf-idf) score. Thereby, the following two options were considered:

1. **Topic-specific Relevance**
   In this option, the tf-idf relevance score will be calculated on the basis of the topic-specific corpus. This includes words that are only derived from articles within the dataset. This option can be implemented using the Sklearn TDIDF method on the dataset content.

2. **Global Relevance**
   The other option, includes a global tf-idf score, based on the corpus of the whole Blendle archive. This measure can be extracted using a "more like this" Elasticsearch query on the Blendle archive.

Since the topic-specific relevance is only trained on the corpus of the dataset, it will be capable of assessing relevance within the topic in more detail than global relevance could. Since this study aims to increase viewpoint diversity within a certain topic, topic-specific relevance has been chosen to be most suitable. Moreover, this matches the approach using framing elements; Baden and Springer describe how frames can be seen as a concrete contextualisation accounting for a specific object or situation. In contrast, they propose interpretative repertoires as generalised points of view [5].

4.5.3. RERANKING RECOMMENDATIONS

Finally, the information from the global diversity measure and the relevance score are used to generate the recommendation for a given article. To do so, the Maximal Marginal Relevance (MMR) algorithm is used. The algorithm works as follows:

1. **Initialise algorithm**
   To initialise the algorithm, two steps have to be taken. First, an article needs to be selected from the dataset for which the recommendations are reranked. This article is added to the set R that includes all articles that are selected by the algorithm so far. Secondly, the number of required recommendations s needs to be set.

2. **Add next article sequentially**
   Afterwards, s new articles are chosen sequentially. Each time, the article with the

maximum marginal relevance compared to articles in the set R is chosen. This value includes a threshold between the relevance score and the diversity score of the new article compared to all other articles in the set. The threshold is represented by the variable λ, which can have a value between 0 and 1. A value of 1 implies an recommendation list that is purely based on relevance, while a value of 0 implies a list that is only based on the diversity measure. During the offline evaluation, the effect of this parameter will be evaluated. These insights will be used to determine a suitable value for λ during the online evaluation.

The MMR algorithm can be efficiently described by the following equation:

$$MMR \equiv max_{i \in R \setminus S}[\lambda(Rel(i) - (1-\lambda) max_{j \in S}(1 - Div(i||j))] \tag{4.8}$$

5

DATA

5.1. INTRODUCTION

After the methodology has been determined, multiple data sets are needed to optimise the model variables and evaluate the performances of the model. For all evaluation studies, including the rule-based suggestion mining evaluation, offline evaluation and online evaluation specific data sets were required. Since the Blendle archive comprises over five million articles from over 150 different sources, a retrieval procedure including strict requirements was developed to obtain the desired data sets. This chapter starts by describing this procedure, including general requirements, that was used to compose each dataset. Afterwards, the data set specific requirements are presented for all evaluation studies, including the rule-based suggestion mining evaluation, offline evaluation and online evaluation. Additionally, analytical information is presented for each dataset.

5.2. DATASET RETRIEVAL

As described before, the data sets are composed from an archive of more than 5 million articles. This section describes the general procedure that was used to compose a dataset. Additionally, the general requirements that were applied on each data set are presented.

5.2.1. PROCEDURE

The procedure to create a data set can be described by the following two steps:

1. **Compose Elasticsearch Query**
   The Blendle archive is stored in Elasticsearch and can thus be searched using a call to the API service. Therefore, a search query needs to be composed which filters the archive based on specific information. This information can include simple features, such as the number of words or the name of a publisher, but more advanced features are added by Blendle as well. Examples include, NER-tags, extracted channels and complexity scores.

2. **Manual Quality Check**
   After the articles are retrieved using the Elasticsearch query, a final manual check needs to be performed. The purpose of this step is twofold. First, since the archive comprises such a large amount of articles, the Elasticsearch response includes many unrelated articles. Therefore, each article is manually checked to meet the predefined requirements and to match the topic of the data set. Secondly, some articles contain content at the end of the article that is unrelated to the topic. Regularly, a reference to another article is included or the curriculum vitae of the author is presented. Since the proposed method considerably depends on the structure of an article, unrelated content at the end of the article is removed manually.

5.2.2. GENERAL REQUIREMENTS

Independent of the application within this study, general requirements have been set up that should apply to every dataset. Table 5.1 provides an overview of these requirements. First, all articles in the dataset should be available on the Blendle platform. Generally, this includes a minimum quality check by Blendle, to remove undesirable content, such as the weather or short actualities. Additionally, since the proposed method heavily relies on the structure of the article, a filter on a minimum number of words and a minimum number of paragraphs is applied. Finally, the article type must be 'opinion piece' or 'background analysis'. This is based on the outcome of the focus group session, in which the structure of these type of articles is put forward as main heuristic for finding framing aspects.

Filter	Value
Available on Blendle.com*	True
Minimum article length	450
Minimum number of paragraphs	5
Article type	Opinion Piece or Background Analysis

*Note that all articles from publishers are picked by Blendle to be available on the platform

Table 5.1: Overview of general data set requirements

Besides these general rules, specific requirements were set up for each dataset. These can be found in the section corresponding to each dataset.

5.3. RULE-BASED SUGGESTION MINING EVALUATION

This section describes the details of the dataset that was created for the evaluation of the rule-based suggestion mining method. First, the requirements are described that were applied in combination with the general requirements, described in the previous section, to compose the dataset. Afterwards, some properties of the final data set are presented.

5.3.1. REQUIREMENTS

To obtain a suitable data set for the evaluation of the rule-based suggestion mining, additional filters were applied in the Elasticsearch query. As described in the literature

study, suggestion mining involves the task of retrieving sentences that contain advice, tips, warnings and recommendations from the opinionated text [56]. Based on this definition, it is assumed that the dataset needs to have a large share of opinionated articles to be able to verify the performances of the proposed method. Additionally, to ensure the method is generally applicable, no restrictions on the topic of the included articles were applied.

Based on these insights, the following three conditions were introduced in the Elasticsearch query. First, a filter was included that selects articles for which the entity 'column' or 'opinion journalism' was extracted. Additionally, the dataset was filtered on publishers that generally have a notable share of opinionated news articles. This includes all national newspapers that are available on Blendle. Finally, to limit the results, only articles that were published in 2019 and 2020 were selected. An overview of these conditions can be found in table 5.2.

Filter	**Value**
Entity Match	column, opinion journalism
Publisher	De volkskrant, Het Algemeen Dagblad, De Standaard (belgium), Trouw, Het Parool.
Article publish date	from 2019

Table 5.2: Specific filters in Elasticsearch Query to retrieve rule-based suggestion mining evaluation data set

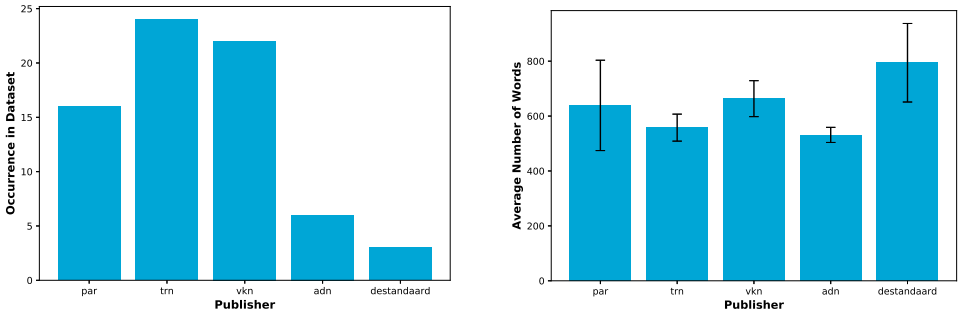
5.3.2. DATA PROPERTIES

Table 5.3 provides an overview of the most relevant properties of the final dataset. The dataset consist of 71 articles from five different publishers. The average number of words is 654, the minimum 450 and the maximum 1169.

Property	**Value**
Number of articles	71
Unique Publishers	5
Average number of words	654
Minimum number of words	450
Maximum number of words	1169

Table 5.3: Properties of data set that was used to evaluate the rule-based suggestion mining for news article content

Additionally, Figure 5.1a provides an overview of number of articles per publisher. It can be seen that multiple different publishers are represented in the dataset, but the number of articles per publisher is not balanced. Additionally, Figure 5.1b illustrates the average number of words per publisher, including the standard deviation.



(a) Number of articles per publisher

(b) Average number of words per publisher

Figure 5.1: Number of articles and average number of words per publisher

5.4. OFFLINE AND ONLINE EVALUATION

5

For the offline and online evaluation the same data sets were used, such that the results of the offline evaluation, such as the optimal model variables, can be used in the online evaluation. Section 5.4.1 provides a description and justification for each topic that was selected. Afterwards, some relevant data properties for each dataset are presented in section 5.4.2.

5.4.1. TOPICS AND TOPIC-SPECIFIC REQUIREMENTS

Since the online experiment was conducted live on the Blendle platform, the recommendation had to be applied on ongoing topics. Therefore, in the week before the experiment, four data sets on ongoing topics were created. Additionally, the topics were chosen to be disputed, such that different viewpoint on the issues were present in the news. Eventually, the following topics were selected: Black Lives Matter movement, Corona Virus, Big-tech and U.S elections.

As described before, the articles are retrieved from an Elasticsearch database using a search query. Three types of additional filters were applied. First, the 'content match' filter makes sure that at least one of the key-words or key-phrases is included in the text. Secondly, the 'content match not' filter does the opposite: excluding articles based on key-words or key-phrases. Thirdly, a filter on the publish date of an article enables the selection of specific time frames for the data set. Below the conditions are described for each data set:

BLACK LIVES MATTER

The first topic is about the Black Lives Matter movement, caused by the death of George Floyd by an U.S. police officer. A publish date filter has been applied such that the articles do not cover the actualities about the first days of the protest, but rather the fundamental discussion about racism that followed. Additionally, articles with the keywords 'belastingdienst' and 'corona' were removed. These cover a certain subtopic about particular issues or events and are therefore considered as unrelated.

Filter	Value
Content Match	'black lives matter', 'racisme debat', 'blm-demonstraties', 'George Floyd', 'racisme-debat'
Content Match Not	'belastingdienst', 'corona'
Article publish date	From 15-06-2020

Table 5.4: Overview of Elasticsearch filters for the topic of Black Lives Matter

CORONA VIRUS

Secondly, the topic of the corona crisis was covered. Since the corona crisis includes many different phases, only the first part of summer of 2020 was covered. During this phase, the restrictions were relatively stable and a few subtopics, such as obligation of mouth maskers and state aid to companies, were under a stable amount of discussion.

Filter	Value
Content Match	'corona', 'covid-19', 'mondkapjes', 'mondkapje', 'mondmasker', 'mondkapjesplicht', 'coronatest', 'coronatesters', 'rivm', 'virus', 'viroloog', 'golf', 'topviroloog', 'uitbraak', 'uitbraken', 'coronaregels', 'versoeplingen', 'staatssteun', 'vaccin'
Article publish date	From 01-06-2020

Table 5.5: Overview of Elasticsearch filters for the topic of corona

U.S. ELECTIONS

Thirdly, the upcoming United Stated presidential elections were chosen as suitable topic. During the weeks before the online experiment, an increasing number of articles was published on the topic. To exclude irrelevant articles about the preliminary election and focus on the presidential battle between Joe Biden and Donald Trump, a filter was applied to only include articles after June first 2020. Additionally, filters based on related key-words and key-phrases to these issues were applied.

Filter	Value
Content Match	'Donald Trump'
Content Match	'presidentsverkiezingen', 'Verkiezingen', 'verkiezingsstrijd', 'campagne', 'verkiezingscampagne', 'Joe biden'
Article publish date	From 01-06-2020

Table 5.6: Overview of Elasticsearch filters for the topic of the U.S. Elections

BIG TECH

The final topic includes the power- and privacy issues of big tech companies, such as Amazon, Google, Apple and Facebook. This topic became actual again after the CEO's of these companies were heard by the House Judiciary Antitrust subcommittee at July the 29th. In contrast to the others, this topic is covered at a lower but constant rate for a couple of years. Therefore, articles since 2018 are included in the dataset. The topic is mainly about the market dominance of these tech giants and privacy issues related to their systems. Filters based on related key-words and key-phrases to these issues are applied.

Filter	Value
Content Match	'macht', 'machtig', privacy', 'data', 'privacy-schandaal', 'privacy-onderzoek'
Content Match	'big tech', 'tech-bedrijven', techbedrijven'
Article publish date	From 2018

Table 5.7: Overview of Elasticsearch filters for the topic of big tech

5.4.2. DATA PROPERTIES

Table 5.8 provides an overview of some properties of each data set that was used during the offline and online evaluation. As can be seen, the size of data set varies between 42 and 69 articles from 6 to 10 different publishers. The average number of words over all data sets includes 700, the minimum 458 and the maximum 2080. Additionally, some properties related to the presentation of article on the Blendle platform were listed. It can be seen that ratio of articles that include thumbnail image highly depends on the topic. For the Black Lives Matter and Corona data set, the majority of the articles are presented with a thumbnail image, while for the U.S. Elections and Big Tech data set the opposite holds. Also, the number of custom titles from the editorial team and the average title length differs considerable per topic.

Property	Black Lives Matter	Corona	U.S. Elections	Big Tech
Number of articles	69	52	42	51
Unique Publisher	10	7	6	10
Average number of words	697	608	744	761
Minimum number of words	464	465	458	458
Maximum number of words	1164	954	1217	2080
Articles with thumbnail	39	27	20	17
Articles with editorial title	1	4	8	10
Average title length	6.3	5.2	9.6	8.1

Table 5.8: Properties of the four data sets that were used for offline and online evaluation

The word distribution is illustrated in Figure 5.2, including the average and standard deviation for each topic. It can be seen that the average number of words per article slightly differs per data set but is far above the minimum of 450. Additionally, the standard deviation is considerable large for the Big Tech topic compared with, for example, corona. Although the average of the Big Tech dataset is comparable with the data set on the U.S. elections, the first thus includes some outliers in terms of the number of words.

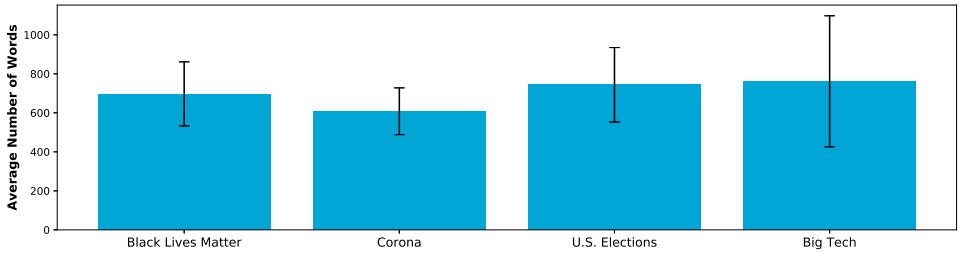


Figure 5.2: Average number of words and standard deviation per data set

Finally, Figure 5.3 provides an overview of the relative proportion of each publisher in a data set for all topics. It can be seen that four publishers are represented in all data sets: De Volkskrant, De Standaard, Trouw and Het Algemeen Dagblad. Furthermore, De Volkskrant is the most prominent publisher in all data sets, except for the topic of the U.S. Elections. The inclusion of other, less frequent, publishers varies per topic.

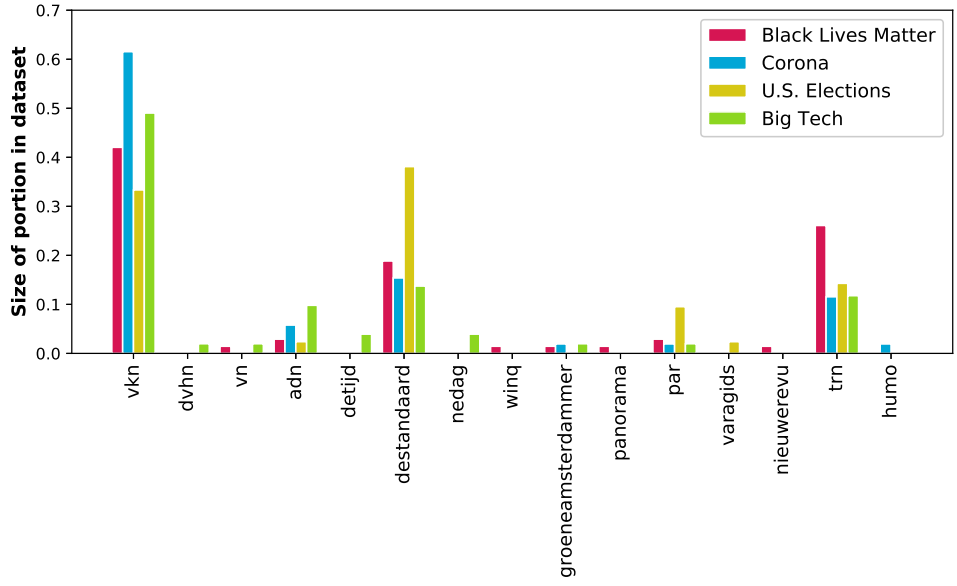


Figure 5.3: Publisher ratio per data set

6

EVALUATION OF RULE-BASED SUGGESTION MINING

6.1. INTRODUCTION

As introduced in section 4.3.3, the methods that are proposed in the research domain of rule-based suggestion mining can provide an interesting starting point in the retrieval of metadata related to the treatment recommendations of the main frame of an article. The task described as suggestion mining involves the extraction of sentences from an opinionated text that include advice, tips, warnings and recommendations [56]. In the light of this study, it is assumed that output of these methods contain information that can be related to treatment recommendations.

Although most research in this domain focuses on domain specific content, mainly product reviews, some general approaches were approached using a rule-based method [57, 56]. However, these general approaches have not been evaluated on news article content. Therefore, the performances of these methods for the application scenario of study, news article content, are evaluated. Additionally, the results can be used to optimise the extraction parameters of this method for the application scenario of this study.

The chapter starts with a description of the experimental setup in section 6.2. Afterwards, the results are presented in section 6.3. Finally, the chapter is concluded in section 6.4.

6.2. EXPERIMENTAL SETUP

As described in the introduction, general suggestion mining approaches using rule-based method are evaluated on news article content. For that purpose, a ground truth data set on suggestions in news article content was needed. As described in the literature study in section 2.5, however, only a limited number of data sets on suggestion mining exists and moreover, all these data sets are based on product reviews. Therefore a ground truth data set was composed using a *crowdsourcing task*, in which participants were asked to annotate suggestions in news article content from the data set that was described in

section 5.3. Afterwards, the performances of the general applicable rules for suggestion mining were evaluated on this dataset. Section 6.2.1 elaborates on the crowdsourcing web application that was developed to obtain the ground truth data set. Afterwards, the general rules for suggestion mining from literature that were evaluated are presented in section 6.2.2. Finally, section 6.2.3 discusses the participants of the crowdsourcing task.

6.2.1. CROWDSOURCING WEB APPLICATION

Although multiple widely-used crowdsourcing platforms exist, the services do not support the specific task of annotating suggestion in news article content. Additionally, the publication of content from Blendle on third-party applications could be in conflict with publisher's terms and conditions. Therefore, a web application for this specific task was developed for internal usage. During four iterations, the application was tested with potential annotators and improved based on their feedback. An example window of the final version is illustrated in Figure 6.1.

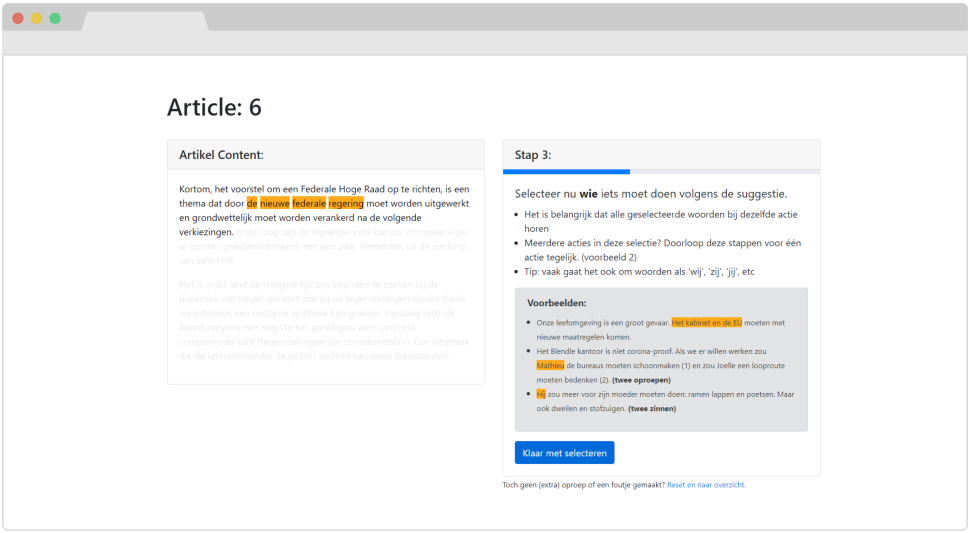


Figure 6.1: View on crowdsourcing web application

In the crowdsourcing web application, users are asked to annotate suggestions in the last two paragraphs of 10 articles. Thereby, the application guarantees that every article gets annotated by at least three unique users. First, a participant needs to accept the terms and conditions related to the task. The terms and conditions are inline with the prescription from the TU Delft and have been approved by the ethical committee. Afterwards, an article is annotated using the following procedure:

ARTICLE ANNOTATION PROCEDURE

Depending on the number of suggestions per article, the procedure is completed multiple times. In every step, the content that needs to be annotated is shown in a window on the left. On the right, instructions and examples are given. The annotation steps include:

1. **Is there any suggestion?**
   First, a user is asked if there is any suggestion present in the text. If not, the next article is loaded. If yes, the actual annotation procedure is started.

2. **Select sentences of suggestion**
   Secondly, an annotator needs to select the sentences that together contain a single suggestion. In most cases, this comprises one sentence. However, some edge cases exist in which a single suggestion spans multiple sentences.

3. **Select the subject of the suggestion**
   Thirdly, the sentences of interest are highlighted and the users are asked to select the subject of suggestion by asking: "who should do something according to the suggestion"?. For example in the sentence "we should fight climate change", "we" is the subject of the suggestion.

4. **Select what is suggested**
   After the subject of the suggestion is selected, the tool would like to know which word or clauses tell what should be done by the subject according to the suggestion. This can also be called the object of the suggestion. For example in the sentence "we should fight climate change", "should fight climate change" tells what should be done.

5. **Selected verbs in suggestion**
   Fifthly, the annotators are asked to select the verbs in the previous selection of what should be done. For example in the sentence "we should fight climate change", "should" and "fight" are the verbs.

6. **Disambiguation**
   As a final step for each suggestion, the users are asked to add any disambiguation for the selected object and subject. Because in many cases, the subject and words refer to something in the context of the full article or context of the whole topic that is being addressed. For example in the sentence "we should fight climate change", "we" refers to society as whole.

7. **Overview**
   Lastly, all previously annotated suggestions related to the article that are being annotated are listed. Users have the option to delete any mistakes, to start the procedure for another suggestion in the article, or continue to the next article.

A visual overview of all steps can be found in appendix A.

6.2.2. RULES

As described in section 2.5, most rule-based suggestion mining approaches focus on a specific domain of content, such as product reviews. However, some studies aim to combine the domain specific rules of different to a set of general applicable rules [57, 56]. Although these have not yet been tested in the context of news articles, it is assumed that these general rules are most applicable for the purpose of this study. Therefore,

Rule	Related work	Patterns
Modal verbs	[66]	Modal verb (can, would, could, must, may, shall, might, should, will, ought to) + Base form of verb (VB)
Imperative detection	[57]	Verb in front of noun
Keywords + Syntactic Patterns	[53]	Base form of verb (VB) + 'option', 'stop' + gerund or present participle (VBG)
Keywords + Phrases	[66] [87] [53] [27]	suggest, recommend, hopefully, go for, request, it would be nice, adding, should come with, should be able, could come with, I need, we need, needs to, need to, would like to, would love to, I wish, I hope, hopefully, if only, would be better if, would that, I can has, do want, should, would like if, can't believe didn't, don't believe didn't

Table 6.1: Overview of rules included in rule-based suggestion mining

the performances of these rules were evaluated. An overview of these rules, including reference to the original work, is provided in table 6.1.

The first type of rules detects suggestion by the sequence of a modal verb followed by the base form of an other verb. Examples of this construction include "must stop" or "should prevent". The second rule aims to find imperative constructions, such as "stop spending money on the EU", by evaluating the presence of any verb in front of any word that could be the subject. Thereby, interrogative sentences are ignored. The third rule is based on a matching on both lexical and syntactic clues. Concretely, this comprises two rules: the presence of any base verb followed by the word "option" and the presence of the word "stop" followed by a verb gerund or present participle form. Lastly, the fourth rule uses the separate sentences from the preprocessing directly and performs a matching with a list of keywords and phrases.

6.2.3. PARTICIPANTS

It was expected that the annotation task requires no specific skills or experience apart from a degree in high school. Therefore, the task was originally published to all Blendle employees. Afterwards, to gain some additional annotation participants, some fellow students were asked to participate as well. No reward was involved in the task.

6.3. RESULTS

After a period of two weeks, 71 articles were annotated by at least 3 unique annotators. As described in the experimental setup, the annotation task involved, apart from selecting sentences the contain a suggestion, more advanced assignments, such as the decomposition of the suggestion into subject and object. However, due to time-constraint, meth-

ods that are capable of using this advanced information were eventually not considered. Therefore, this section focuses only on the results related to task of labeling sentences that contain a suggestion.

First, the inter-rater agreement is calculated using the *Krippendorff's alpha coefficient*. Afterwards, the results for each rule presented in Table 6.1 will be discussed. As described in section 6.2, an evaluation study was performed to assess the performance of rule-based suggestion mining on news articles. Thereby, the general applicable rules from previous research on suggestion mining were used. A crowdsourcing task was published to create a ground truth data set on suggestion in the concluding paragraphs of opinionated news articles.

For the evaluation, the choice has been made to only focus on the question if a sentence contains any suggestion. Although more detailed information was gathered during the crowdsourcing task, such as the subject and object of the suggestion, this information will not be used. This is mainly a result of the choice to not use the role extraction information in the pipeline related to the treatment recommendation, described in section 4.3.3. Additionally, it has been decided to focus on the implementation of the full pipeline rather than the optimisation of each individual function.

6.3.1. INTER-RATER AGREEMENT

To provide an insight in the difficulty of the crowdsourcing task and the reliability of the results, the inter-rater agreement reliability was calculated. Since the task includes multiple annotators who annotate only a subset of all articles, the Krippendorff's alpha coefficient was used [41]. By using the ratio between the observed disagreement D_0 and the disagreement expected by change D_e, the coefficient corrects for the change of a certain outcome. The following equation describes the Krippendorff's alpha coefficient.

$$\alpha = 1 - \frac{D_0}{D_e} \tag{6.1}$$

The calculate the coefficient for the suggestion mining crowdsourcing task, a canonical form matrix of the response data was created. Every sentence of every article obtained a unique column in the matrix, such that the response of a unique annotator could be described by a single row. Since the task was performed by 26 unique annotators on 547 sentences, the matrix was of size 26×547. Each entry could contain an 0, 1 or *. In case of a zero, the annotator did not label the sentence as a suggestion. In case of a one, the annotator did label the sentence as suggestion. In case of a *, the annotator did not see the sentence at all (the subset of annotated articles by the annotator did not contain the article of the sentence).

The analyses returned a Krippendorff's alpha coefficient of **0.49**. According to the standards for the coefficient, this can be seen as a moderate agreement level.

6.3.2. PERFORMANCES OF RULES

Table 6.2 provides an overview of the results of each rule that was described in the experimental setup. Thereby, the number of true positives, the cases in which the sentence was labelled to contain a suggestion by both the method and at least two participants of the crowdsourcing task, and false positives, in which the sentence was labelled by the

method but by less than two participants of the crowdsourcing task, are included in the table.

Rule	Pattern	True Positives	False Positives
Modal	can	6	29
	would	1	8
	could	1	2
	must	6	9
	may	1	6
	shall	0	0
	might	0	1
	should	3	6
	will	3	31
	ought to	0	0
Keywords	we need	2	0
	needs to	1	3
	should to	0	8
	suggest to	2	0
	ability to to	2	0
	request	1	0
	if only	1	0
	if hopefully	1	0
	i want	1	0
Imperative		8	18
Syntactic		0	0

Table 6.2: Overview of performance general rules for suggestion mining

Besides, the true positives and false positives presented for each rule in Table 6.2, the number of false negatives was calculated to be 17. This implies a precision of 0.18, recall of 0.62 and a F1 score of 0.28.

6.4. CONCLUSION

These results can be compared with the study of Negi et al. who assess the performances of same set of rules on available datasets, such as product reviews in tweets or fora [57]. These results are presented in Table 6.3.

Generally, it can be seen that the evaluation of these rules on news article content obtains comparable results to the study of Negi et al. [57] on product related content. The general rules are thus applicable in the news domain. Compared to other approaches for suggestion mining, however, the rule-based approach still yields very limited performances. Due to time consideration, it was decided not to implement any other method based on the crowdsourced data set. For the purpose of the offline and online evaluation, however, it has been decided to remove the worse performing clues from the method. This includes rules for which the ratio between true positives and false positives is very

Data Set	Precision	Recall	F1
Electronics Reviews	0.229	0.660	0.340
Hotel Reviews	0.196	0.517	0.285
Travel Discussion 2	0.312	0.378	0.342
Microsoft Tweets	0.207	0.756	0.325
New Tweets	0.200	0.398	0.266
Suggestion Forum	0.461	0.879	0.605

Table 6.3: Results of rule-based suggestion mining on available data sets by Negi et al. [57]

low. From the modal verb patterns, can, will, might and would were removed. Additionally, should was removed as keyword. Finally, the syntactic rule was removed in total, since no overlap was found in the dataset.

7

Offline Evaluation

7.1. Introduction

As final step before running the online evaluation, related to the main research question of this study, the offline evaluation is performed. This experiment provides an answer to the last sub research question; After the methodology was set up, the data sets were created and the variables corresponding to the rule-based suggestion mining were optimised, the offline study was used to verify the model's capability of enhancing the viewpoint diversity of recommendation lists, corresponding to the following research question:

RQ 4: *Is the proposed method capable of increasing the viewpoint diversity of recommendation lists, according to a metric from literature?*

Additionally, the effect of the general model variables needs to be assessed, such that they can be optimised for the online evaluation. Section 7.2 describes the experimental setup of the offline evaluation. In section 7.3, the results of the offline study are represented. Finally, the chapter is concluded in section 7.4.

7.2. Experimental Setup

This section describes the experimental setup of the offline evaluation. First, the evaluation procedure is described. Afterwards, the model variables, metric and baseline are described, respectively.

7.2.1. Procedure

As described before, the offline and online evaluation are performed on the same data sets. These include four data sets, each on a specific topic that are both actual and disputed during the online experiment. Further details and the justification for the choice of these data sets, can be found in section 5.4.

For each data set, the performance of the model is assessed for different variations of the model variables, such that the optimal values can be used in the online evaluation. The procedure can be described by the following four-step procedure:

1. **Enrich dataset**
   First, all articles in the dataset are enriched according to the four framing aspects of the definition of Entman: the problem definition, causal attribution, moral evaluation and treatment recommendations [20]. Details of this process can be found in section 4.3.

2. **Generate Diversity and Relevance Matrices**
   Secondly, all combinations of two articles are compared, based on the enrichments. Therefore, a distance function was implemented for each framing aspect. This distance function includes a measure for the dissimilarity of two articles based on the framing aspects. Details of each framing function can be found in section 4.4. Afterwards, this information is used to generate the global diversity matrix, as described in section 4.5. Finally, since the MMR-algorithm reranks a list based on a linear combination between diversity and relevance, the TF-IDF relevance matrix is calculated, including a relevance score for each two-article combination.

3. **Optimise model variables and evaluate performances using cross-validation**
   Thirdly, for each article i in the dataset, a set of s recommendations is calculated by reranking the remainder articles in the dataset. To prevent overfitting, this calculation is being performed using cross-validation. For that purpose, the dataset is split into k distinct sets. For every set, the following steps are taken:

   (a) **Grid search of model variables on training set**
   The training set contains the $k-1$ subsets of articles. Using a grid search, the optimal combination of the model variables are obtained for the training set. An overview of the model variables can be found in section 7.2.2.

   (b) **Evaluation on test set**
   After the variables are trained on the $k-1$ subsets, the model is evaluated on the test set for different values of lambda. As described before, for each article in the test set, a set of s recommendations is calculated by reranking the remainder articles in the total dataset. The evaluation metric and other measures variables are described in section 7.2.3.

   During the offline evaluation, the effect of different values of the cross-validation variables k and the size of the recommendation list s are assessed as well. Due to time-constraints, however, these values will not be optimised using, for example, nested cross-validation. For the size of the recommendation list s, three different values are chosen. The settings $s = 3$ and $s = 6$ are chosen to roughly match the current sizes of the recommendations on the Blendle platform. Additionally, $s = 9$ is chosen to be relatively larger than the other values but still relevant for real-life applications. For the cross-validation parameters k, three different values are evaluated. $k = 10$ is included because this is regularly seen as the standard setting for the cross-validation setting. Additionally, $k = 5$ and $k = 20$ are chosen to be a

factor two different from the regular setting and to roughly match the values for the size of the recommendation list. For example, $k = 20$ will result in group size of about 3, considering the average data set size that is being used.

4. **Combine results**
   Eventually, the results of all k cross-validation parts are combined to be analysed.

7.2.2. MODEL VARIABLES

Table 7.1 provides an overview of the model variables that are optimised during the offline evaluation. Since the global diversity matrix is composed as the weighted sum over the four distance functions related to each framing function, the first four model variables include the corresponding weight factors. The range of the weight factors are chosen such that each framing aspect contributes always for some part and a single framing function can not be dominant. A step size of 0.1 is assumed to provide enough variations for the weight factors. Additionally, a model variable for the taxonomy level weight is included: equal weights for each taxonomy level or ascending weights for each taxonomy level ([1,..,5]). As described in section 4.4.2 and 4.4.3, ascending weight factors would enable a larger contribution of higher-level categories to the diversity measure. However, higher-level taxonomy categories could also be too detailed in the light of this study. Therefore, equal weight model variables have been included. Besides, a model variable for the number of paragraphs that is assumed to be introductory and a model variable for the number of paragraphs that is assumed to be concluding have been introduced. Both variables can be either 1 or 2. Finally, the value of λ in the MMR algorithm is optimised as well. Thereby, a step size of 0.1 is assumed to provide enough variation to the variable.

Variable	**Values**
Weight Framing function - Problem Definition	[0.1, 0.2, 0.3, 0,4]*
Weight Framing function - Causal Attribution	[0.1, 0.2, 0.3, 0,4]*
Weight Framing function - Moral Evaluation	[0.1, 0.2, 0.3, 0,4]*
Weight Framing function - Treatment Recommendation	[0,1, 0.2, 0.3, 0,4]*
Taxonomy level weight	[equal, ascending]
Number of introducing paragraphs	[1, 2]
Number of concluding paragraphs	[1, 2]
λ	[0.0, 0.1, ..., 0.9]

*Note that all framing function weight factors should sum up to 1

Table 7.1: Overview of possible values of model variables

7.2.3. VIEWPOINT DIVERSITY METRIC AND ADDITIONAL MEASUREMENTS

During the offline evaluation, the performance of the model for different variables is assessed. Most importantly, a metric from literature is used to asses the model's capability of enhancing viewpoint diversity. However, the effect of the model on other data set properties will be interesting as well. For example, how will the publisher ration be af-

fected by the model? First, the viewpoint diversity metric is discussed. Afterwards, these additional measurements are described.

VIEWPOINT DIVERSITY METRIC

To assess the performance of the viewpoint diversification method, a metric is needed. As described in the literature study in chapter 2, only little research has been done on metrics for viewpoint diversity in recommender systems. Moreover, current measures for viewpoint diversity have been criticized for lacking theoretical foundation [32, 67]. However, the development of a novel metric for viewpoint diversity is not a focus of this thesis.

The most comparable work is the work by Tintarev et al., who aims to increase viewpoint diversity in a recommender system by using article data from Blendle. [82]. Therefore, the model is evaluated according to the metric presented in their work. This includes the *Intra-list Diversity* and is used in several studies on diversity in recommender systems [94, 86, 93, 82]. ILD is defined as follows:

$$Diversity = \frac{\sum_{i=1}^{n} \sum_{j=i+1}^{n} Distance(D_i, Dj)}{n \times (n-1)/2} \tag{7.1}$$

Thus, the viewpoint diversity is defined as the average distance between all pairs of articles i and j, such that $i \neq j$. Thereby, the distance between a pair is defined by the channels and topics (LDA) as derived from the enrichment methods:

$$Distance(i, j) = 0.5 \times Distance_{Channels} + 0.5 \times Distance_{LDA} \tag{7.2}$$

Here, the channel-distance is calculated using the cosine distance, whereas the LDA-distance is computed using the Kullback-Leibler divergence. The channels and LDA distribution can be directly retrieved from Blendle enrichment data.

ADDITIONAL MEASUREMENTS

As described before, besides the viewpoint diversity metric, the effect of the diversification model on other properties is evaluated as well. Below, the measurements are listed:

- **Relevance**
  As described in the literature study in section 2.2, diversification approaches generally aim to increase diversity while maintaining the relevance of the recommendations. Therefore, the TF-IDF relevance is measured for each recommendation list, such that the effect of the viewpoint diversification method on this feature can be evaluated.

- **Kendalls Tau**
  To be able to analyse if the proposed method is capable of providing different recommendations sets compared to the baseline method, the *Kendall rank correlation coefficient* τ is calculated [39]. This coefficient provides a measure for the similarity of two ranks of data.

- **Average number of words**
  Additionally, the average number of words for each article in the recommendation lists is measured. Although there is no reason to suppose that an 800 words article is preferable above a 600 words article, it is assumed that a lower limit on the number of words can be seen as an heuristic for a minimum quality of news articles. Blendle itself assumes that an article has reasonable quality when it has more than 450 words. Therefore, the method should be able to maintain a reasonable average number of words.

- **Publisher Ratio**
  Finally, the publisher ratio is measured for the recommendation lists. Potentially, this could provide insights on the effect of the content diversity on the source diversity. As described in the literature study, source diversity is the most common way to enhance viewpoint diversity. However, scholars argue that viewpoint diversity can only by achieved by fostering content diversity [89].

7.2.4. BASELINE

To assess if the proposed diversification method is able to increase the viewpoint diversity based on the presented metric, a baseline is needed. The baseline should be chosen such that the method has minimum effects on the recommendation in terms of viewpoint diversity. Two options have been considered:

A **Full Relevance MMR** In this option, the baseline is implemented using a MMR setting of $\lambda = 1$, such that the recommendations are purely ranked on the TF-IDF relevance.

B **Random Order** In this option, the baseline is implemented using a random rerank of the recommendation list.

Eventually, the first option including a MMR setting of $\lambda = 1$ has been chosen the most suitable for the evaluation. This choice is partly based on the online evaluation; In this experiment, one group of Blendle users receives viewpoint diverse recommendations, while the control group receives recommendations from a baseline. These recommendations are provided below an article in a section called "read further on this topic". A full description of the online evaluation can be found in section 8.2 It is assumed that a baseline that is based on full relevance enhances a better user experience for this section, compared to a random list. Additionally, the most comparable work of Tintarev et al. implements the baseline using a full relevance rerank as well [82].

7.3. RESULTS

In this section, the results of the offline evaluation are presented. First, the optimal model variables are described for each topic, cross-validation parameters k and size of the recommendation list s. Afterwards, the performances of the optimal model for different values of λ are described, including the viewpoint diversity and relevance scores. Finally, the effect of the model on other properties of the recommendation list are presented, including the Kendall rank correlation coefficient, the average number of words and the publisher ratio.

7.3.1. OPTIMAL MODEL VARIABLES

In Table 7.2, the optimal model variables are presented for each topic, cross-validation parameters k and size of the recommendation list s. The optimal setting involves the combination for which most articles in the data set obtained the maximum viewpoint diversity score. On average, the optimal setting yields the best performances for 84% of all articles in the data set. Therefore, only these settings are presented in Table 7.2.

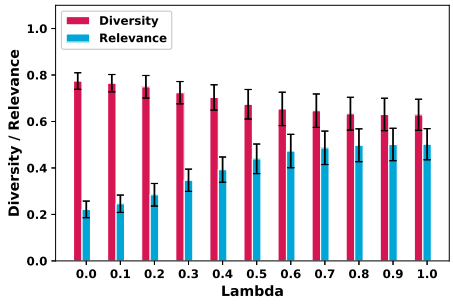
From the results, it can be seen that the optimal number of introductory and concluding paragraphs is very dependent on the topic. Besides a few exceptions, these numbers are relatively stable across different values for k and s. The same holds for the category weight setting. In contrast, the general weights are also relatively stable within one topic but a general pattern across topics can also be found. Generally, it can be seen that the first and third weight factor obtain smaller values than the second and fourth weight factor. Thereby, the second factor is most stable with only one exception across all combinations of modal variables and topics. Finally, $\lambda = 0$ which implies full diversity yields the best performances in terms of viewpoint diversity.

Topic	k	s	intro. par	concl. par	general weight	category weight	λ
Black Lives Matter	5	3	2	1	[0.2, 0.4, 0.1, 0.3]	eq	0
	10	3	2	1	[0.2, 0.4, 0.1, 0.3]	eq	0
	20	3	2	1	[0.2, 0.4, 0.1, 0.3]	eq	0
	10	6	2	1	[0.4, 0.4, 0.1, 0.1]	eq	0
	10	9	2	1	[0.4, 0.4, 0.1, 0.1]	eq	0
Corona	5	3	2	1	[0.1, 0.4, 0.1, 0.4]	eq	0
	10	3	2	1	[0.1, 0.4, 0.1, 0.4]	eq	0
	20	3	2	1	[0.1, 0.4, 0.1, 0.4]	eq	0
	10	6	2	1	[0.1, 0.3, 0.2, 0.4]	eq	0
	10	9	2	1	[0.1, 0.3, 0.2, 0.4]	asc	0
U.S. Elections	5	3	1	2	[0.1, 0.4, 0.1, 0.4]	eq	0
	10	3	1	2	[0.1, 0.4, 0.1, 0.4]	eq	0
	20	3	1	2	[0.1, 0.4, 0.1, 0.4]	eq	0
	10	6	2	2	[0.1, 0.4, 0.1, 0.4]	asc	0
	10	9	1	2	[0.1, 0.4, 0.1, 0.4]	asc	0
Big Tech	5	3	1	2	[0.2, 0.4, 0.1, 0.3]	asc	0
	10	3	1	2	[0.2, 0.4, 0.1, 0.3]	asc	0
	20	3	1	2	[0.2, 0.4, 0.1, 0.3]	asc	0
	10	6	2	1	[0.2, 0.4, 0.1, 0.3]	eq	0
	10	9	2	2	[0.2, 0.4, 0.2, 0.2]	asc	0

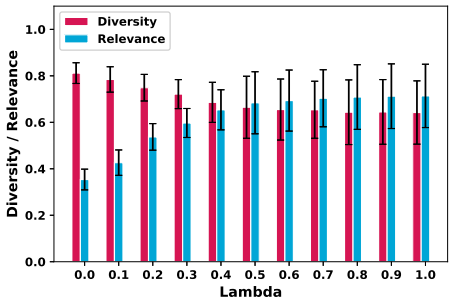
Table 7.2: Best performing setting of model variables for each data set, cross-validation k and size of recommendation list s

7.3.2. VIEWPOINT DIVERSITY AND RELEVANCE

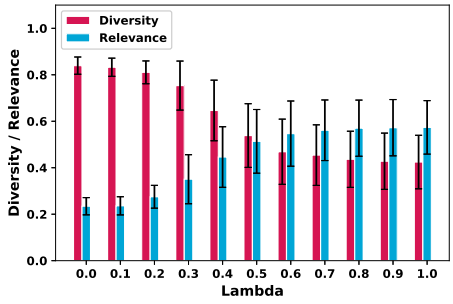
For the optimal setting of the model variables, the performance of the model in terms of the viewpoint diversity and relevance score for different values of λ are illustrated in Figure 7.1. As described before, a λ setting of 1.0 refers to the baseline and reranks a list purely on relevance, while a λ setting of 0.0 implies full diversity. All values in between represent a linear combination of relevance and diversity. The red bars represent the results of the viewpoint diversity metric, while the blue bars represent the relevance scores. The black shapes describe the standard deviation of these measures for different recommendation lists. For each data set, the cross-validation variable was fixed to $k = 10$ and the list size to $s = 3$.



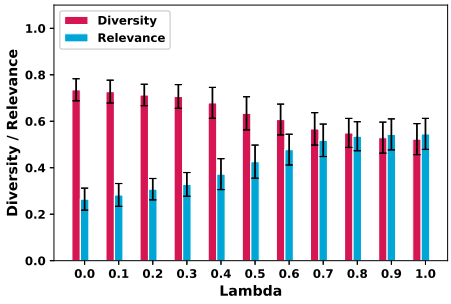
(a) Topic: Black Lives Matter



(b) Topic: Corona Virus



(c) Topic: U.S. Elections



(d) Topic: Big Tech

Figure 7.1: Diversity and relevance scores for different values of λ per topic.

Across all topics, it can be seen that the proposed diversification method is capable of increasing the viewpoint diversity of recommendation lists. According to the metric, the viewpoint diversity increases on average from 0.55 to 0.79. Besides, the average relevance decreases from 0.58 to 0.27. Furthermore, it can be observed that the gradient of both the increase of diversity and decrease of relevance over all values of lambda is not constant. For each topic, a phase can be identified in which the gradient is at it's maximum. For example. for the topic of corona this includes the phase between $\lambda = 0.0$ and $\lambda = 0.4$. Lastly, it can be observed that the standard deviation increases for larger values of λ.

INFLUENCE OF CROSS-VALIDATION VARIABLE k

For each topic, the influence of the cross-validation variable k on the results has been evaluated. The results for the Black Lives Matter topic are illustrated in Figure 7.2. Again, the red bars represent the viewpoint diversity scores and the blue bars represent the relevance score. However, in this figure the shades of each color represent different values for cross-validation variable. The lightest color refers to five cross-rounds, one step darker represents 10 cross-rounds and the darkest variant stands for 20 cross-rounds.

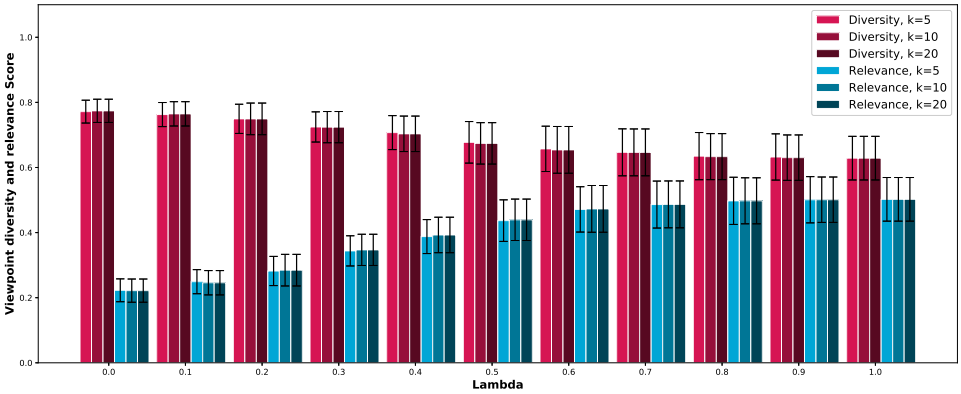


Figure 7.2: Relevance and viewpoint diversity scores across different values of λ and k for the Black Lives Matter topic and $s = 3$

Generally, it can be be seen that there is no or only little influence of the cross-validation variable k on the results. There is almost no or no consistent variation of the heights of different shades of each bar observable. Likewise, the standard deviation shows no remarkable fluctuation. The same result applies for other topics as well.

INFLUENCE OF RECOMMENDATION LIST SIZE s

Likewise, the influence of the recommendation list size s has been evaluated. The results for the Black Lives Matter topic are illustrated in Figure 7.3. Again, the red bars represent the viewpoint diversity scores and the blue bars represent the relevance score. However, in this figure the shades of each color represent different values for the size of the recommendation list. The lightest color refers to 3 recommendations, one step darker represents 6 recommendations and the darkest variant stands for 9 recommendations.

In general, it can be observed that for larger values of lambda larger recommendation list yield both higher viewpoint diversity scores and smaller relevance scores. Additionally, it can be observed that the standard deviation decreases for larger sizes of the recommendation list. These results are similar across other topics as well.

7.3.3. KENDALLS TAU

As described before, the Kendall rank correlation coefficient τ is measured to provide insight in the similarity of different recommendation lists. The coefficient is bounded between -1 and 1, where -1 represents highly different and 1 represents totally overlap-

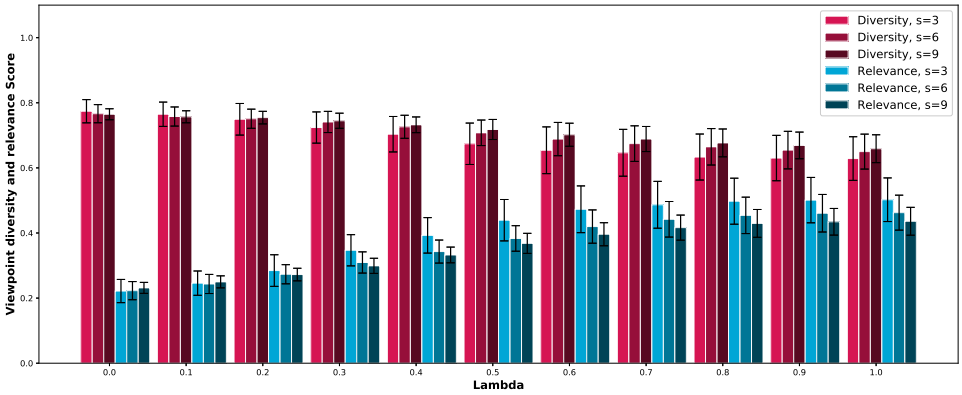


Figure 7.3: Overview of relevance and viewpoint diversity scores across different values of λ and s, for the Black Lives Matter topic and $k = 10$

ping ranks. The objective of this measurement is to assess whether the proposed diversification method is capable of providing different recommendation lists compared to the baseline. Therefore, the coefficient has been calculated for the combination of the baseline ($\lambda = 0$) with each other value of lambda ($\lambda = [0.1, 0.2, ..., 1.0]$).

Figure 7.4 and 7.5 provide an overview of the average Kendall rank correlation coefficient and standard deviation across all topics. In the figures, each topic is represented by a specific color. Additionally, the shades of the color in Figure 7.4 represent different values of the cross-validation variable k, whereas the shades in Figure 7.5 represent different values of the recommendation list size s.

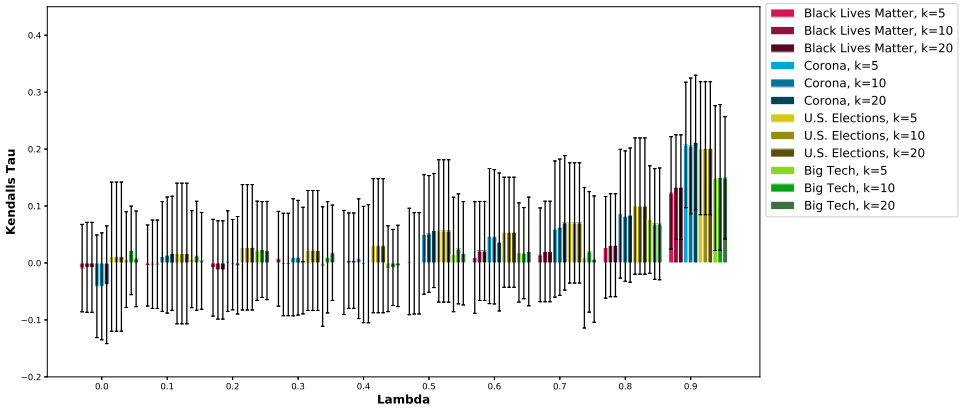


Figure 7.4: Kendalls Tau score relative to the baseline for recommendation lists of size $s = 3$ and different values of λ, topic and k

From the results of Figure 7.4, it can be observed that reranking the set of recommendations based on viewpoint diversity results in different recommendation lists compared to the baseline. Additionally, it can be seen that the coefficient decreases for smaller val-

ues of λ. This drop, however, seems to be bounded around $\tau = 0$ for decreasing values of λ. Besides, it can be observed that the influence of the cross-validation variable k is very small. Also, no consistent pattern can be observed in the variation of the results due to k.

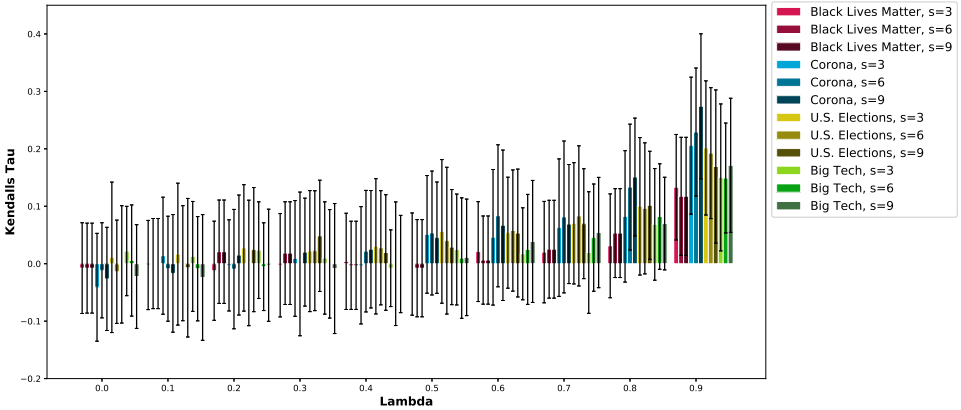


Figure 7.5: Kendalls Tau score relative to the baseline for different values of λ, topic and s. The cross-validation fold is fixed to $k = 10$

Comparable results can be obtained from Figure 7.5. However, the influence of the size of the recommendations lists s appears to be larger than the influence of k. In both figures, it can be observed that the standard deviation is not largely effected by the value of λ, k or s.

7.3.4. AVERAGE NUMBER OF WORDS

Additionally, the average number of words of the recommendation lists and the standard deviation across different recommendation lists are measured. Figure 7.6 and 7.7 provide an overview of the results for all topics. In the figures, each topic is represented by a specific color. Similar to the previous section, the shades of the color in Figure 7.6 represent different values of the cross-validation variable k, whereas the shades in Figure 7.7 represent different values of the recommendation list size s.

From Figure 7.6 and 7.7, no consistent pattern can be observed in the average number of words for different values of λ. In case of the Black Lives Matter and Big Tech topic, the number increases for larger values of λ, for the topic of the U.S. Elections the average decreases and for the topic of Corona the average remains stable. Similar to Kendall Tau, the cross-fold has only minor influence, both in terms of the average in the standard deviation. The influence of the size of the recommendation list s appears to be larger. Although no consistent pattern can be observed for the average across topic, the standard deviation increases for larger values of λ.

7.3.5. PUBLISHER RATIO

To provide an insight on the publisher ratio of the recommendation lists for different values of λ, different results can be analysed. Figure 7.8 illustrates the results for three

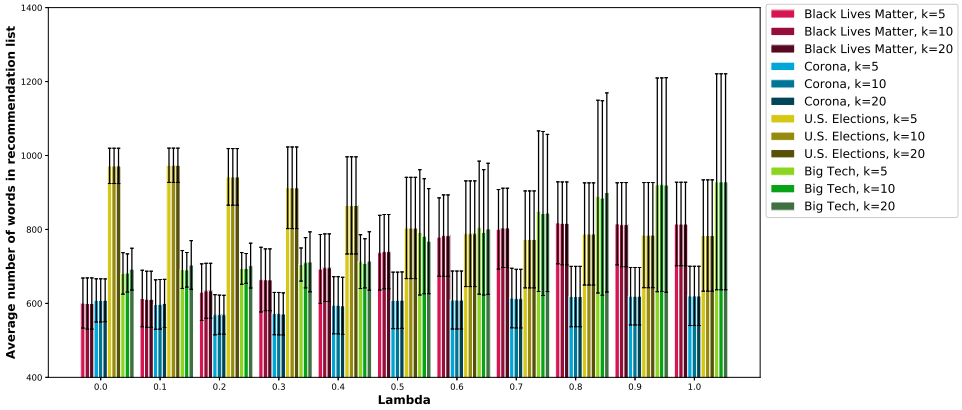


Figure 7.6: Overview of average number of words for different values of λ, topic and k. List size is fixed to $s = 3$

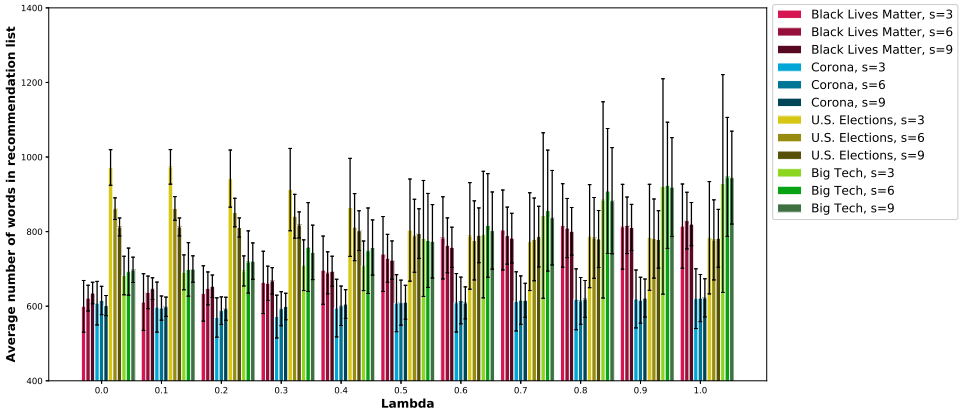


Figure 7.7: Overview of average number of words for different values of λ, topic and s. Cross-validation fold is set to $k = 10$

different options of analysing the data for the Black Lives Matter topic. First, the number of articles for each publisher can be counted. Secondly, the rank of each article recommendation can be taken into account, such that higher ranks have a larger influence to the score. Finally, the count of the number of articles was normalised compared to the input ratio.

For farther analyses, the third representation has been chosen to be most suitable. Firstly, because the differences between the count and rank are very small for every data set. Secondly because the normalisation relative to input ratio ensures that the results purely present the effects of the model.

Figure 7.9 illustrates the results of each topic for different values of λ. The cross-validation variable was fixed to $k = 10$ and the list size to $s = 3$.

From Figure 7.9, multiple results can be observed. Firstly, it can be seen that the

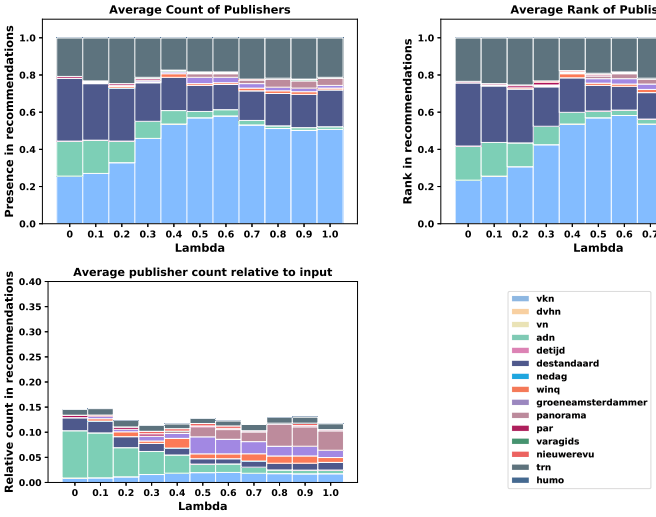


Figure 7.8: Three different options for analysing publisher ratio results applied for the Black Lives Matter topic

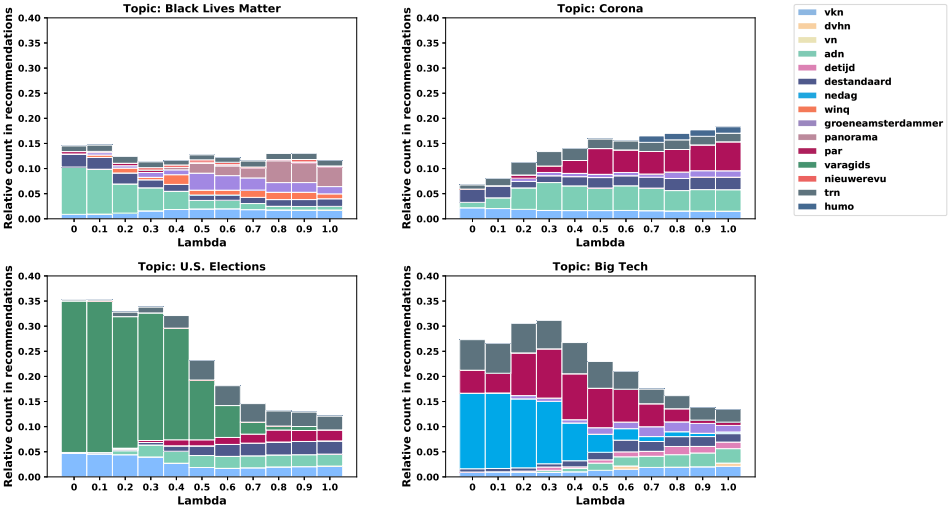


Figure 7.9: Average count of publishers in recommendations lists normalised for the ratio of the input for all four topics, $k = 10$ and $s = 3$

number of represented publishers increases for larger values of λ; The number of different publishers in the baseline recommendation list is thus larger than the number that is represented in the full diversity recommendation list. This applies for every topic. The same results can also be observed when the rank ratio is not normalised relative to the input ratio. Secondly, it can be observed that the diversification method significantly influences the publisher ratio. For small values of λ some publisher get remarkably am-

plified, while others are completely excluded. This effect can primarily by seen in the third and fourth sub-figure. The corona topic seems to be the only exception.

INFLUENCE OF CROSS-VALIDATION VARIABLE k

For each topic, the influence of the cross-validation variable k on the publisher ratio has been evaluated as well. The results for the Black Lives Matter topic are illustrated in Figure 7.2.

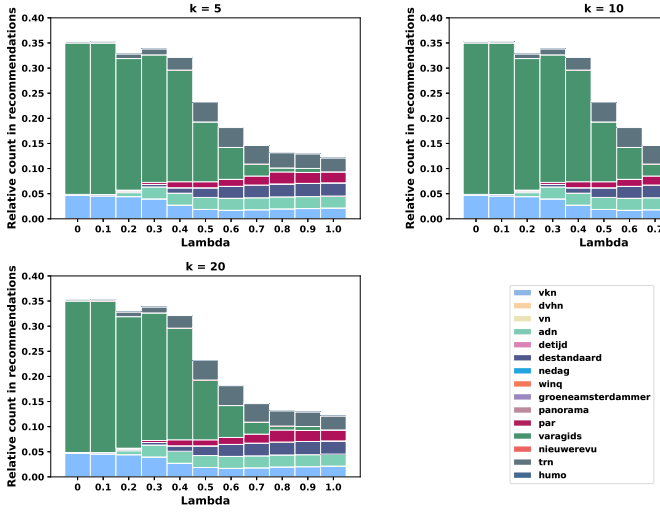


Figure 7.10: Influence of cross-validation variable k on publisher ratio for the Black Lives Matter topic and $s = 3$

Similar to the other results, the influence of the cross-validation variable seems to be very small; Only minor changes to the ratio can be observed. The same result can be observed for other topics as well.

INFLUENCE OF RECOMMENDATION LIST SIZE s

Additionally, the influence of the size of the recommendation list s on the results has been evaluated. Since the publisher ration is highly different per topic, Figure 7.11 includes the results for all four topics for the maximum list size of $s = 9$.

Compared to the results with a smaller lost size of $s = 3$, illustrated in the beginning of this section in Figure 7.9, a larger list size moderates both effects for decreasing values of λ. Thus, both the drop in the number of included publisher is smaller compared to a smaller list size and the amplification and moderation effects to a publisher a reduced. Similar effects can be obtained to a lesser extent for $s = 6$.

7.4. CONCLUSION

As described in the introduction of the section, the offline evaluation has been performed to be able to answer the final sub research question, before the online evaluation

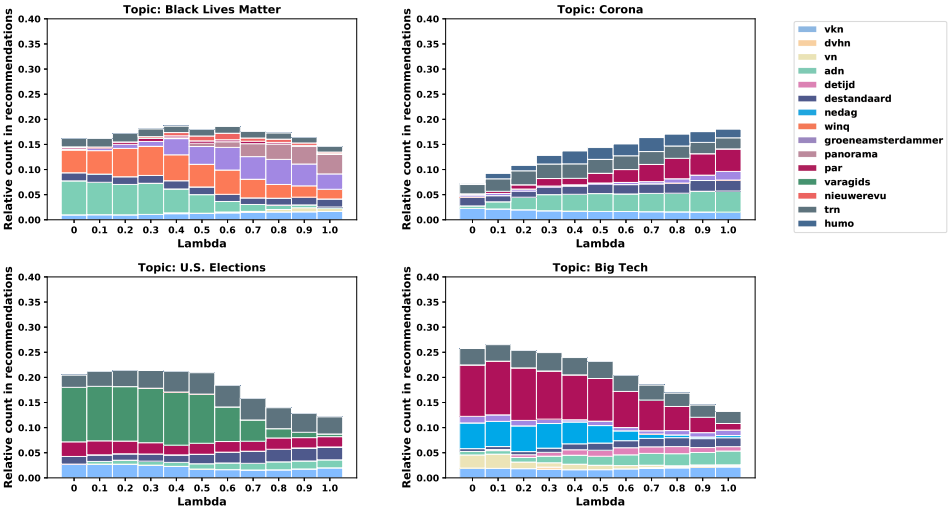


Figure 7.11: Average count of publishers in recommendations lists normalised for the ratio of the input for all four topics, $k = 10$ and $s = 9$

related to the main research question can take place. This involves the following question:

RQ 4: *Is the proposed method capable of increasing the viewpoint diversity of recommendation lists, according to a metric from literature?*

Based on the results, it can be concluded that the method is capable of increasing the viewpoint diversity of recommendations lists, according to the metric defined by Tintarev et al. [82]. For decreasing values of λ, which implies a larger contribution of diversity in the MMR-algorithm, the viewpoint diversity scores increase among all evaluated topics, values for the size of the recommendation list s and values for the cross-validation parameter k. Additionally, from the results of the Kendall rank correlation coefficient, it can be observed that proposed diversification is capable of providing different recommendation lists using reranking; For decreasing values of λ, the coefficient is reduced as well. In the light of the online evaluation, this is an important finding. During the experiment, one group of users receives recommendations from the baseline method, while the other group of users receives the recommendations with increased viewpoint diversity. Decreasing values of the rank correlation coefficient thus ensure that both groups receive different lists.

Besides, from the results, it can be seen that the diversification method has an influence on the average number of words of recommendation lists. The influence seems, however, to be different per topic. Therefore, this property of the recommendation set will be evaluated during the online experiment as well. For example, the length of an article can have an influence on the reading behaviour; Short articles are more likely to be finished, while longer articles demand more time from a user. It is, however, not as-

sumed that the difference in average word length has other implications, for example, for the quality of the recommendation list. All lists are far above the minimum of 450 words that Blendle uses as lower bound for the article length.

Apart from these main findings, some additional conclusions can be drawn. First, the exact choice of the value of λ seems to have an implication on different properties of the recommendation lists. The results of the viewpoint diversity scores indicate that the increase of diversity is not constant for decreasing values of λ. Rather, a phase can be identified per topic for which the growth is at its maximum. Also, the average rank correlation coefficient slows down for decreasing values of λ and seems to bounded around $\tau = 0$. Finally, decreasing values of λ result in the exclusion of certain publishers for all topics. Also, for the majority of the topics some publishers get amplified remarkably. In the experiment setup of the online evaluation, described in section 8.2, the choice for the value of λ for the online experiment is justified based on these findings.

Finally, some conclusions can be drawn related to the cross-validation parameter k and the recommendation list size s. First, across all results the influence of the cross-validation parameter k has found to be very small. Therefore, this parameter will be fixed to the default of $k = 10$ for the online evaluation. The list size s, however, does have a slightly larger influence on the results. In particular, increased list sizes decrease the standard deviation for multiple list properties. The only exception seems to be the Kendall rank correlation coefficient. Also, larger list sizes moderate both the exclusion of publishers and the amplification of certain other publishers for smaller values of λ. In the online experiment, however, the list size was bounded to $s = 3$ due to limitations by Blendle and their contracted publishers. This is described in more detail in section 8.2.

8

Online Evaluation

8.1. Introduction

At this point, all sub research questions are answered. First, in contrast to the definitions in current diversification approaches, diversity in media is understood as viewpoint diversity. Thereby, a conceptualisation of this definition through framing theory was seen as most suitable for a novel diversification method for news media. Secondly, the definition of a frame by Entman, in which four framing aspects are described, was considered most suitable for this study [20]. Afterwards, a method was implemented to extract metadata related to these framing aspects using nlp-toolkits. Thirdly, this metadata was combined to a viewpoint diversity measure, which was used in a MMR algorithm, such that recommendations could be reranked based on this viewpoint diversity measure. Finally, the method was evaluated during an offline evaluation. Among other results, the evaluation indicates that the method is capable of increasing the viewpoint diversity of recommendation lists using reranking. Thus, all steps are taken to perform the final evaluation related to the main research question of this study:

Main RQ: *How is reading behaviour affected by viewpoint diverse news recommendations and how they are presented?*

To be able to answer the main research question, an online study is conducted on the Blendle platform. In this study, users receive recommendations from either the novel proposed method or the baseline method. Thereby, multiple measurements are conducted to analyse the difference in reading behaviour between users groups. This chapter starts with a detailed description of experimental setup in section 8.2. Afterwards, the results are presented in section 8.3. Finally, the chapter is concluded in section 8.4.

8.2. Experimental Setup

As described before, the online study is performed to assess how the proposed diversification method affects the reading behaviour of users. For that purpose, an experiment is

set up on the Blendle platform, which includes an iOS-, Android- and web-application.

On all these devices, the today page presents a daily news article selection to a user. This page can be seen as the central functionality of the service. As described in the introduction in section 1.2, the daily article selection for a user is composed out of two sections. First, the editorial selection involves the five 'must reads' of the day, according to the editorial team. Secondly, the editorial selection is complemented by a more personalised set of articles, generated by the recommender system of Blendle. Figure 8.1 provides an overview of this today page, including the two sections.

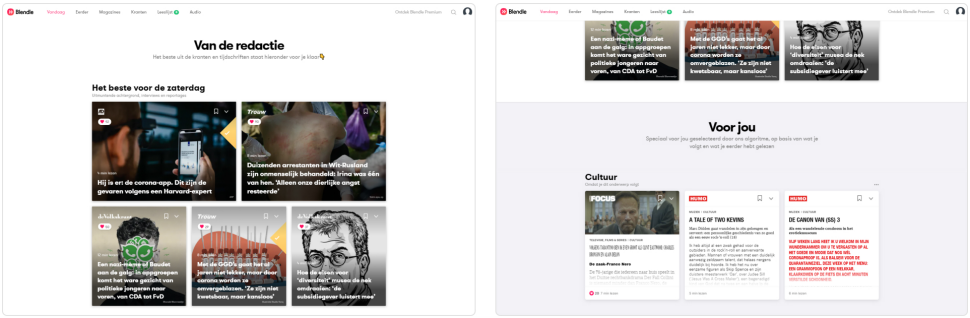


Figure 8.1: Two scroll positions of the today page of Blendle, including the editorial selection in the left image and the personalised selection in the right image

When an article is selected by the user, a reading page opens in which the content of the article is shown. In the current setup of Blendle, the space below the content of the article on this reading page is used to repeat the not already read articles from the today page. In this way, users are not forced to go back to the today page to continue reading in Blendle.

For the purpose of the experiment, however, a new functionality is implemented which enables the presentation of recommendations on the same topic below the selected article. This section is called "read further on this topic" and can include three different recommendations to the original article. Figure 8.2 provides an example of both the normal functionality, on the left, and the new functionality related to this experiment, on the right.

8.2.1. CONTRACTUAL LIMITATIONS

Blendle has contractual agreements with all publishers which content is included in the platform. In the light of this study, the only relevant agreement involves the maximal number of DPG Media articles that can be included per user per day. This involves articles from 13 publisher that are owned by DPG Media. For the purpose of this study, a maximum number of four DPG Media articles can be included in the recommendations per day.

The influence of these restrictions have been briefly analysed using a sample run. In this sample run, the relevance and diversity scores of recommendation lists that were produced without any restrictions were compared with recommendations from the restricted method. The results indicated that if at least two out of three articles in the

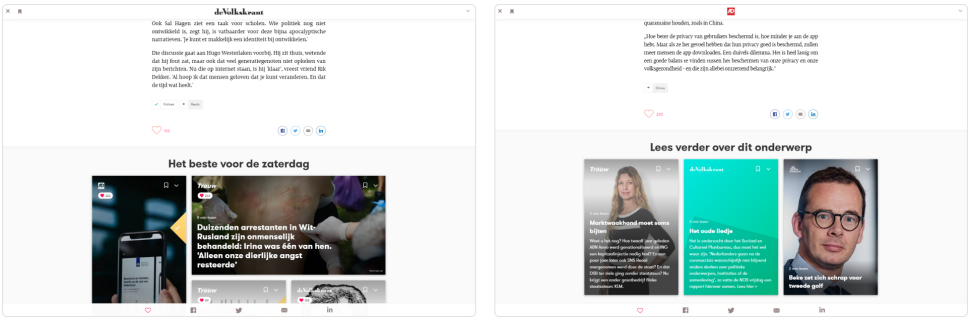


Figure 8.2: In the normal situation, on the left, the articles of the today page are repeated below the article content. In the new functionality related to the online experiment, three recommendations on the same topic are provided

recommendation set can be DPG Media articles, the influence of the restriction on the diversity and relevance score is limited. Based on this conclusion, it was decided to produce recommendations for no more than two articles every day. Then, each article can include two DPG Media articles, such that the total number of articles from DPG Media is up to four in total.

In addition, due to the restriction of two articles per day for which recommendations can be provided, it was decided to only provide recommendations below articles that were selected by the editorial team. First, because these articles are generally about ongoing topics. Secondly because the editorial selected articles are by far the most read articles, obtaining around 1500 reads per article every day.

8.2.2. BASELINE

Since the goal of the online evaluation was to evaluate the influence of the proposed method on the reading behaviour, the diversification method needed to be compared against some baseline method. Therefore, two groups of users were created, such that one group received recommendation from the diversification method, while the other group received recommendation from the baseline method. As described in the offline evaluation, the baseline was implemented using a MMR that was fully based on relevance ($\lambda = 1.0$). A justification of this choice can be found in section 7.2.4.

8.2.3. USERS

At the time of the experiment, Blendle was used by around 60k paying customers. From these users, a selection was made to only include users who clicked in the last 14 days before the experiment at least 4 times an article that was provided below the content of another article. These users were assumed to most likely see and use the new recommendation functionality. In total, 2076 users were selected. Afterwards, these users were split randomly into two equal-size groups of 1038 users. One group, thus, will receive diversified recommendation, while the other group receives articles that are only relevant to the original article.

8.2.4. DATA

Since the proposed method is only capable of diversification within the topic, multiple data sets on a specific topic were prepared preliminary to the experiment. As described in chapter 5, these data sets were chosen to be disputed topics that were under discussion in the news during the time of the experiment. These data sets were used in the offline evaluation as well and are described in more detail in chapter 5.

Before the start of the experiment, all four data sets were enriched to retrieve the metadata related to the four framing functions, as described in section 4.3. By extracting this information before the start of the experiment, the daily process of calculating the recommendation lists with new articles could be done more quickly.

8.2.5. MODEL VARIABLES

The goal of the offline evaluation was to both verify the model's capability of enhancing the viewpoint diversity of recommendation lists and to analyse the influence of different setups of the model variables on the viewpoint diversity metric, relevance and other properties, such as the publisher ratio. By using the same data sets for the online and offline evaluation, the model variables could be optimised for the online evaluation.

The model variables, including the number of introduction paragraphs, the number of concluding paragraphs, the weight factors of the viewpoint diversity measure and the category weight, were fixed per data set to the value that was found optimal during the offline evaluation. Additionally, the values for a list size of $s = 3$ and cross-validation setting of $k = 10$ were chosen. As described above, the new functionality to provide recommendations below an article is implemented by Blendle to provide three recommendations. Due to time restrictions of both this study and Blendle, no other values for s were evaluated. As described in the conclusion of the offline evaluation in section 7.4, the value of the cross-validation setting has no or only a small influence on the results. Therefore, the standard setting of $k = 10$ was chosen.

Lastly, the value of λ needed to be determined for the diversification method. Although it would interesting to evaluate different values of λ on the reading behaviour, only one setting could be chosen due to time considerations. Eventually, it was decided to use $\lambda = 0$ which yielded a maximum viewpoint diversity score for all topics during the offline evaluation. This decision is mainly based on previous experiments with users on the Blendle platform. According to the data team, it has been found difficult to obtain any significant differences in the results between two user groups during multiple experiments in the past years. Therefore, $\lambda = 0$, which yields the maximum difference with the baseline in terms of viewpoint diversity, was found to be the most suitable choice.

Topic	**k**	**s**	**intro. par**	**concl. par**	**general weight**	**category weight**	λ
Black Lives Matter	10	3	2	1	[0.2, 0.4, 0.1, 0.3]	eq	0
Corona	10	3	2	1	[0.1, 0.4, 0.1, 0.4]	eq	0
U.S. Elections	10	3	1	2	[0.1, 0.4, 0.1, 0.4]	eq	0
Big Tech	10	3	1	2	[0.2, 0.4, 0.1, 0.3]	asc	0

Table 8.1: Overview of model variables that were used during the online evaluation for each data set

8.2.6. PROCEDURE

Based on the information provided above, the daily procedure of calculating the recommendations can be described by the following steps:

1. **Select Editorial Articles**
   Every morning at 6:30, the editorial team publishes five 'must reads' on the platform. Directly after publication, these articles were manually checked to match any topic of the four data sets. As described in section 8.2.1, due to contractual limitations between Blendle in DPG Media, a maximum of two of these 'must reads' is selected every day for which recommendations will be provided.

2. **Enrich new articles and calculate diversity and relevance matrices**
   Afterwards, the two selected articles are enriched according to the pipeline described in section 4.3. Based on this metadata and the metadata from all articles in the data set, the diversity and relevance matrices are composed, in which every entry describes the distance or relevance between the new article and an article in the data set. More details of this process can be found in section 4.5.

3. **Calculate recommendation for both user groups**
   Based on the diversity and relevance matrices, the recommendations are calculated using the MMR-algorithm for both the user group that receives the baseline recommendations and the user group that receives the diversified recommendations.

4. **Publish recommendations**
   Eventually, two sets of three recommendations are composed for each article that was selected. Finally, these sets are sent to the Blendle back-end to be published on the platform.

8.2.7. MEASUREMENT OF READING BEHAVIOUR

To analyse the reading behaviour of the two different user groups, specific events can be measured on the Blendle platform, such as which article has been opened by a user or if a user has completely read an article. Based on these available events, multiple measures of the reading behaviour are observed. Below, an overview of both the implicit and explicit measurement methods is included.

IMPLICIT MEASUREMENTS

Three implicit measures of the reading behaviour of users are observed, including the click-through rate per article, the click-through rate per recommendation set and the completion rate of a recommended article. A detailed description of each measure is included below.

1. **Click-through rate per article**
   In this case, the click-through rate is calculated per article. To do so, the number of clicks on an article is divided by the total number of users who have finished one of the original articles for which the article was recommended. Thereby, the completion of an original article is registered using a scroll-position.

2. **Click-through rate per recommendation set**
   Technically, if a user finishes a recommendation, there is the possibility to go back to the original article and click on either of the two other recommendations. It is, however, assumed to be more likely that the user goes back to the today page or selects an article from the today page that is presented below the recommendation. Therefore, it is assumed that a user normally chooses one of three provided recommendations. Therefore, the click-through rate is also calculated over the recommendation set. Thus, the total number of clicks on either of the three articles in the recommendation set, divided by the number of users who have finished the original article for which the recommendation set was presented. Similar to the click-through rate per recommendation, the completion of an original article is registered using a scroll-position.

3. **Completion rate of recommendation**
   The completion rate is implemented as the number of users that completely read the recommended article divided by the number of users who opened the article. The completion rate is assumed to be a measure for the user satisfaction with the recommendations. It can be augmented that short articles are more likely to be completed than long articles. Therefore, the completion rate was also analysed relative to the number of words of an article. Similar to the click-through, the completion of an original article is registered using a scroll-position.

EXPLICIT MEASUREMENTS: HEART RATIO

As described before, Blendle includes a functionality to mark an article as favorite, illustrated by a heart. This heart can be clicked by a user at the end of the article content on a reading page. The measure was implemented as the number of hearts given by the user group, divided by the number of users that completed the article and thus, had the possibility to click the favorite button.

8

8.2.8. INFLUENCE OF RECOMMENDATION PROPERTIES

Additionally, the influence of multiple properties of the individual recommendations or recommendation sets are analysed, including the presentation characteristics, source diversity and article length.

PRESENTATION CHARACTERISTICS

As described in the introduction in chapter 1, it is assumed that the presentation characteristics could have an influence on the reading behaviour of users. On the Blendle platform, a recommendation is presented as rectangular box, including a title, publisher logo and a number of hearts, representing the number of users who selected the article as their favorite. A visual example of the properties can be found in Figure 8.1 and Figure 8.2. The impact of the following aspects on the click-through rate was analysed:

1. **Inclusion of Thumbnail Image**
   The rectangular box can have a thumbnail image as background. It is assumed that recommendation with such a background could gain more attention from users. Therefore, the influence on the click-through rate is assessed.

2. **Title: Length + Editorial title**
   The number of words in the title could potentially have an impact on the click-through rate. Additionally, the original title of an article can be replaced by a custom title of the editorial team. In general, these custom titles are longer and more explanatory than the original titles. Therefore, the difference between articles with and without editorial title will be compared.

3. **Number of Hearts** Finally, the number of hearts that are presented at the recommendation represent the number of users that selected the recommendation as their favorite. It is assumed that a larger number of hearts could have a positive effect on the click-through rate. Therefore, this relationship is analysed.

Since these aspects differ per individual recommendation and not per recommendation set, the click-through rate calculated per recommendation was used.

SOURCE DIVERSITY

During the offline evaluation, it was found that the proposed diversification method affects the source diversity considerable. Based on these results, it was found valuable to assess the influence of the source diversity on the results of the online evaluation as well. Since source diversity only applies to recommendation sets, the impact on the click-through rate calculated per recommendation set was analysed. Thereby, source diversity was conceptualised as the number of different publishers in a recommendation set.

RECOMMENDATION LENGTH

Finally, it is assumed that the length of a recommendation can have an impact on the completion rate. Therefore, the correlation between these two variables is analysed.

8.2.9. STATISTICAL MEASUREMENTS

To evaluate whether the result of the two user groups can be seen as different, the statistical significance is assessed. This section describes which statistical hypotheses tests are used. Additionally, the statistical coefficient that is used to evaluate whether the variables are correlated is described.

SIGNIFICANCE

For all three measures, including the click-through rate, completion rate and hearts ratio, the difference in mean between the two user groups needed to be evaluated. Therefore, a statistical significance is performed to asses whether an observed difference can truly be induced by the method. For all results, the *student t-test*, *Welch's t-test* and *Mann-Whitney U test* are calculated. Each test involves particular assumptions about the experiment and data. Below, these assumptions are described per test.

A **Student t-test**
   The Student t-test is the most commonly applied statistical hypotheses test. The test requires the data of both groups follow a normal distribution and have equal variance. The first assumption is tested using a *Shapiro-Wilk* test on both groups. Additionally, the equal variance is tested using a *Levene's test*.

B **Welch's t-test**
In case the variance of both groups is not equal according to the Levene's test, the Welch's t-test can be used. This test also assumes that the data is normally distributed but does not require equal variance.

C **Mann-Whitney U test**
Finally, if one of the two groups does not appear to be normally distributed according to the *Shapiro-Wilk* test, the Mann-Whitney U test can be used. This test can be used if the dependent variable is either ordinal or continuous, but not normally distributed.

In the results, the validity of each test for the corresponding data is discussed.

CORRELATION

Additionally, the statistical correlation between some properties is measured. For example, it is assumed that the number of words of a recommendation can have an influence on the completion rate. The *Pearson correlation coefficient* is the most commonly used correlation coefficient but assumes a linear correlation between two variables. However, none of the potential variable correlations are assumed to be necessarily linear. Therefore, the *Spearman's rank correlation coefficient* is used instead. This coefficient only assumes a monotonic relationship between two variables and therefore, is assumed to be more suitable.

8.2.10. NORMALITY

To assess if the collected data is normally distributed, the Shapiro–Wilk test will be conducted. For a sample $x_1, ..., x_n$, the test statistic indicates whether the sample came from a normal distribution. The null hypotheses states that the sample came from a normal distribution. In case of a p-value < 0.5, the null hypothesis can be rejected and thus, the data can not be assumed to be normally distributed. In case of a p-value ≥ 0.5, the null hypothesis can not be rejected and the data can be assumed to be normally distributed.

8.2.11. VARIANCE

To assess whether to the variance of two samples is equal, the Levene's test will be conducted. The null hypotheses states that both samples have the same variance. In case of a p-value < 0.5, the null hypothesis can be rejected and thus, it can not be assumed that the samples have equal variance. In case of a p-value ≥ 0.5, the null hypothesis can not be rejected and it can be assumed that the samples have equal variance.

8.3. RESULTS

Eventually, the experiment related to online evaluation ran six days a week for two weeks. Every day, two articles from the editorial 'must read' selection were found to match any of the four topics of the prepared data sets. Therefore, recommendations were provided below 24 articles in total. During the time of the experiment, the topic of corona had gained increasing awareness, related to the fear of a second wave of the spread of the virus. In contrast, the other topics lost considerable attention during the experiment.

This holds especially for the black lives matter topic for which no recommendation could be provided. Therefore, this topic was removed from the analyses. Table 8.2 provides an overview of the number of articles for which recommendations were provided per topic.

Topic	Number of Articles
Total	24
Topic: Black Lives Matter	0
Topic: Corona	18
Topic: U.S. Elections	4
Topic: Big Tech	2

Table 8.2: Overview of number of editorial selected articles per topic for which recommendations were provided during the online experiment

This section describes the results of the experiment on the basis of the three measures of the user behaviour as described in the experimental setup. First, the results corresponding to the click-through rate are analysed. Afterwards, the completion rate of the recommendations is analysed. Thirdly, the results related to heart ratio, the number of users in the experiment that selected a recommendation as their favorite, are presented. Additionally, the influence of multiple data properties on the reading behaviour were analysed.

8.3.1. CLICK-THROUGH RATE

First, the click-through rate of both user groups in the experiment was analysed. As described in the experimental setup in section 8.2, two methods to calculate the click-through rate were considered. In the first option, the click-through rate is calculated per recommendation. Thus, for each recommendation, the clicks on that recommendation are divided by the number of users that completely read the original article. In the second option, the click-through is calculated per recommendation set. Thus, the clicks on all three recommendations that are presented below a certain article are summed and divided by number of users that completely read the original article.

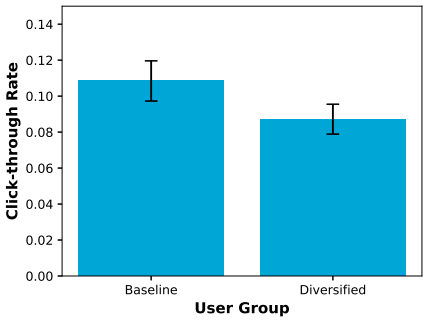
CLICK-THROUGH RATE PER RECOMMENDED ARTICLE

The mean and standard deviation of the click-through rate per recommendation are illustrated in Figure 8.3. Additionally, Table 8.3 provides an overview of the results, including the statistical measures. All calculations were conducted per topic as well and included in Table B.1 and B.2 in Appendix B.

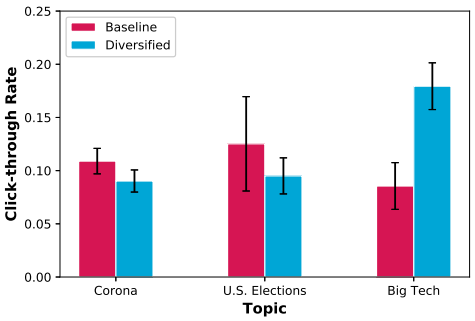
From these results, it can be seen that none of the statistical significance tests yield any p-value ≤ 0.05. Therefore, regardless of which assumptions for statistical tests are met, no significant difference between the two users groups was found related to the click-through rate, calculated per recommended article. The same holds for the results per topic.

CLICK-THROUGH RATE PER RECOMMENDATION SET

The mean and standard deviation of the click-through rate per recommendation set are illustrated in Figure 8.4. Additionally, Table 8.4 provides an overview of the results, in-



(a) Results over all topics



(b) Results per topic

Figure 8.3: Mean and standard error of click-through rate per recommendation both across all topics and per topic

Calculation	**Mean**	**Error**
Baseline	0.11	0.011
Diverse	0.087	0.0083
	Statistic	***p*-value**
Shapiro-Wilk Baseline	0.91	0.0064
Shapiro-Wilk Diverse	0.92	0.01
Levene's	2.2	0.14
Student t-test	1.5	0.13
Welch's t-test	1.5	0.13
Mann-Whitney U	570.0	0.1

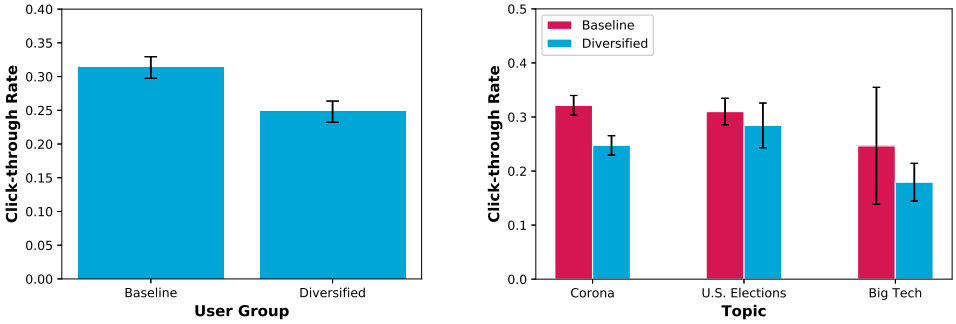
Table 8.3: Overview of the results of multiple statistical measurements for the click-through rate, calculated per recommended article

cluding the statistical measures. All calculations were conducted per topic as well and included in Table B.1 and B.2 in Appendix B.

From the results of the Shapiro-Wilk test, it can be seen that the results of both user groups can be assumed to be normally distributed. Additionally, the Levene's test indicates that the variance of the two user groups can be assumed to be equal. Therefore, the Student t-test can be used to assess whether the mean click-through rate is significantly different between the two user groups. The Student t-test yields a p-value ≤ 0.05 and thus, indicates a significance difference between the mean of the click-through rate per recommendation set of the two user groups of 6.5%. Regarding the result per topic, only the topic of corona yields a p-value ≤ 0.05 for the Mann-Whitney U test. For this topic, the the difference between the baseline and diversified method is 7.4%.

8.3.2. COMPLETION RATE

Secondly, the user behaviour was analysed by means of the completion rate. This involves the ratio between the number of users in the experiment who completely read



(a) Results over all topics

(b) Results per topic

Figure 8.4: Mean and standard error of click-through rate per recommendation set, both across all topics and per topic

Calculation	**Mean**	**Error**
Baseline	0.31	0.016
Diverse	0.25	0.016
	Statistic	***p*-value**
Shapiro-Wilk Baseline	0.96	0.41
Shapiro-Wilk Diverse	0.95	0.27
Levene's	0.048	0.83
Student t-test	2.9	0.0054
Welch's t-test	2.9	0.0054
Mann-Whitney U	160.0	0.004

Table 8.4: Overview of the results of multiple statistical measurements for the click-through rate, calculated per recommendation set

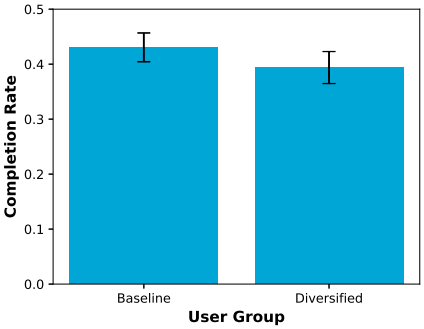
the recommendation, divided by the number of users in the experiment that opened the recommendation.

The mean and standard deviation of the completion rate are illustrated in Figure 8.5. Additionally, Table 8.5 provides an overview of the results, including the statistical measures. All calculations were conducted per topic as well and included in Table B.1 and B.2 in Appendix B.

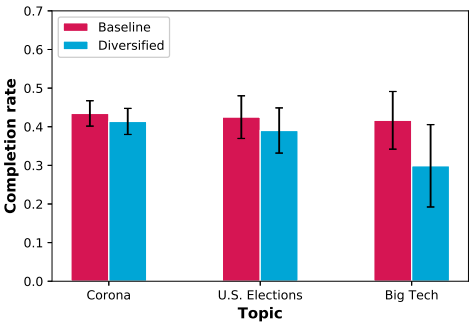
From these results, it can be seen that none of the statistical significance tests yield any p-value ≤ 0.05. Therefore, regardless of which assumptions for statistical tests are met, no significant difference between the two users groups was found related to the completion rate. The same holds for the results per topic.

8.3.3. HEART RATE

Finally, the number of times that a recommendation was selected as favorite, illustrated by a heart in the application, was measured. Since an article can be labelled as favorite at the end of the content, the measure includes the number of hearts divided by the



(a) Results over all topics



(b) Results per topic

Figure 8.5: Mean and standard error of completion rate, both across all topics and per topic

Calculation	**Mean**	**Error**
Baseline	0.43	0.026
Diverse	0.39	0.029
	Statistic	***p*-value**
Shapiro-Wilk Baseline	0.97	0.36
Shapiro-Wilk Diverse	0.98	0.81
Levene's	0.56	0.46
Student t-test	0.94	0.35
Welch's t-test	0.94	0.35
Mann-Whitney U	600.0	0.17

Table 8.5: Overview of the results of multiple statistical measurements for the completion rate

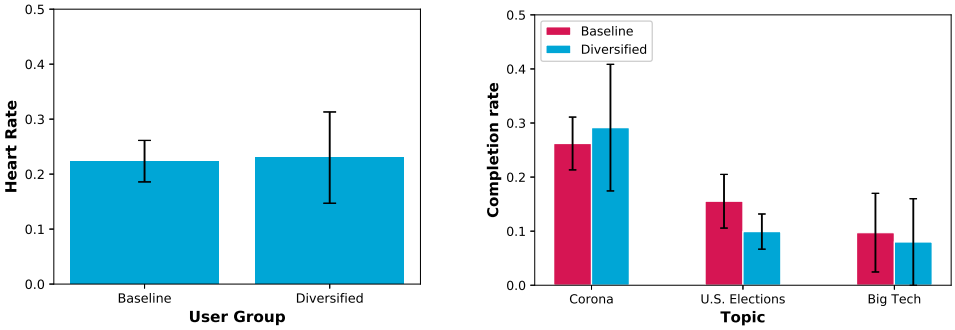
number of users in the experiment that completely read the article.

The mean and standard deviation of the heart rate are illustrated in Figure 8.6. Additionally, Table 8.6 provides an overview of the results, including the statistical measures. All calculations were conducted per topic as well and included in Table B.1 and B.2 in Appendix B.

From these results, it can be seen that none of the statistical significance tests yield any p-value ≤ 0.05. Therefore, regardless of which assumptions for statistical tests are met, no significant difference between the two users groups was found related to the heart rate. The same holds for the results per topic.

8.3.4. INFLUENCE OF DATA PROPERTIES

As described in the experimental setup, some properties of recommendations were assumed to have a potential influence on the reading. First, the influence of the source diversity of the recommendation set on the click-through is analysed. Secondly, the affect of presentation characteristics of the recommendation, such as the inclusion of a thumbnail image, on the click-through rate is assessed. Finally, the impact of the article



(a) Results over all topics

(b) Results per topic

Figure 8.6: Mean and standard error of heart rate per recommendation both across all topics and per topic

Calculation	**Mean**	**Error**
Baseline	0.22	0.038
Diverse	0.23	0.083
	Statistic	***p*-value**
Shapiro-Wilk Baseline	0.78	4.6e-06
Shapiro-Wilk Diverse	0.4	6e-11
Levene's	0.43	0.51
Student t-test	-0.072	0.94
Welch's t-test	-0.071	0.94
Mann-Whitney U	580.0	0.14

Table 8.6: Overview of the results of multiple statistical measurements for the heart rate

length, the number of words of an article, on the completion rate is analysed.

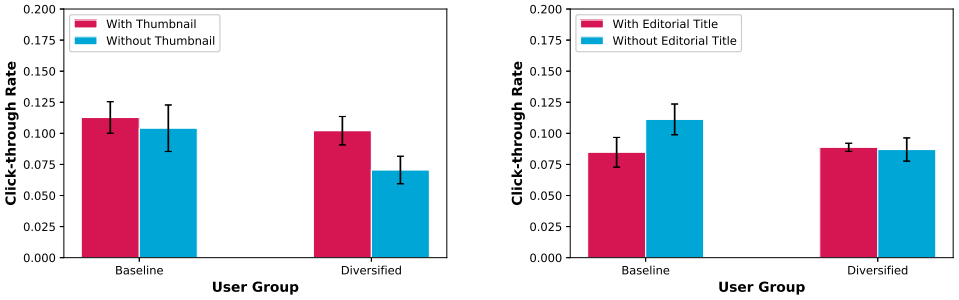
PRESENTATION CHARACTERISTICS

As described in the experimental setup, the influence of multiple properties related to the presentation of a recommended article on the click-through rate were measured. First, the influence of both the presence of a thumbnail image and the presence of a custom editorial title on the click-through rate per recommendation were evaluated.

Figure 8.7 provides an overview of the mean and standard error of the click-through rate per recommendation for recommendation with thumbnail and without thumbnail. Similar, the click-through rate per recommendation for recommendation with editorial and without editorial title are presented. In addition, Table B.3 and Table B.4 in Appendix B include all statistical measurements related to the influence of the thumbnail image and editorial title on the click-through rate.

As Figure 8.7 clearly reveals, no significant influence of the inclusion of a thumbnail image on the click-through rate for baseline users can be found. In contrast the figure suggests a significant difference for diverse users. As Table B.3 shows, the data can not be

assumed to be normally distributed. Therefore, the results of Mann-Whitney test were used. The corresponding p-value of 0.01 confirms the significant difference. Thus, recommendations with a thumbnail are 3.1% more opened than recommendations without a thumbnail for diverse users. Regarding the influence of the inclusion of a editorial title on the click-through rate, it can be seen that none of the statistical significance tests yields any p-value ≤ 0.05. Therefore, regardless of which assumptions for statistical tests are met, no significant influence of the editorial title on neither the baseline and diverse users was found.



(a) Thumbnail (b) Editorial Title

Figure 8.7: Mean and standard error of click-through rate per recommendation for recommendation with or without thumbnail and with or without editorial title

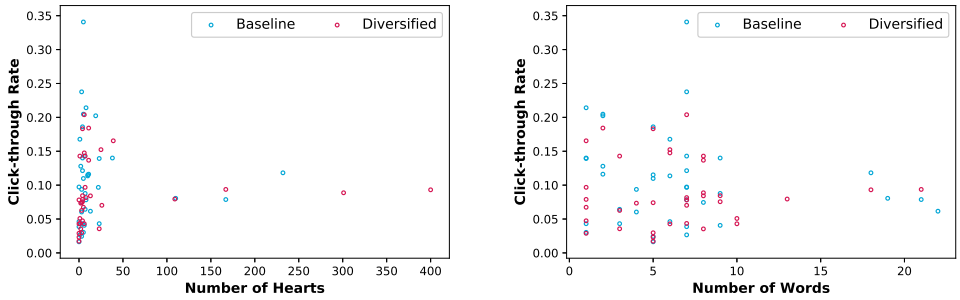
Besides, the correlation of the click-through rate with two continuous variables related to the presentation of a recommendation was assessed. This includes the number of hearts, a representation of the number of users that selected an article as their favorite, and the number of words in the title. Figure 8.8 provides an overview of the relation of both variables with the click-through rate by means of a scatter-plot. Table 8.7 includes the Spearman's rank correlation coefficient for both variables, per user group. From this table, it can be seen that only a significant correlation was found between the number of hearts and the click-through rate for the diverse users. This includes a moderate correlation of 0.57. For the baseline users, no correlation was found.

Property	**Group**	ρ **statistic**	ρ p**-value**
Number of Hearts	Baseline	0.22	0.18
	Diverse	0.57	0.00028
Number of Words	Baseline	-0.12	0.46
	Diverse	0.092	0.59

Table 8.7: Spearman's rank correlation coefficient for correlation between the click-through rate and the number of hearts and between the number of words in the title

SOURCE DIVERSITY

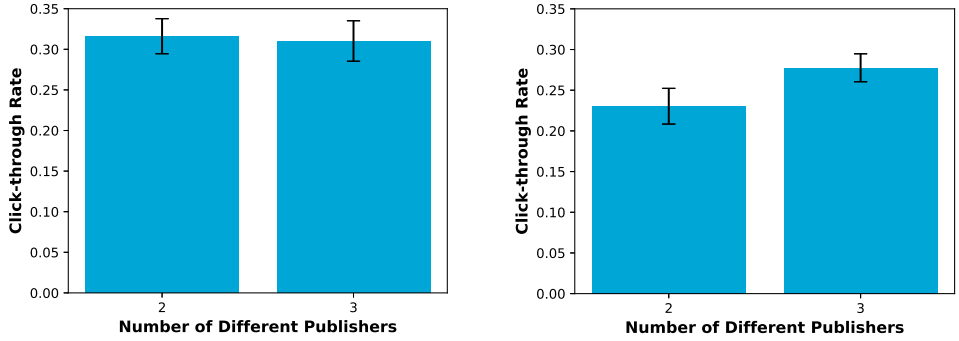
As described in the conclusion of the offline evaluation in section 7.4, higher levels of viewpoint diversity in the offline evaluation turned out to have some remarkable effects



(a) Number of Hearts (b) Number of Words in Title

Figure 8.8: Scatter-plot of relationship between the click-through rate and the number of hearts and between the number of words in the title

on the publisher ratio. Therefore, the effect of the source diversity of a recommendation set on the click-through rate was evaluated. For each recommendation set of three articles, the number of different publishers was calculated. Two categories were found: recommendation sets in which all articles are from a different publisher and sets in which two articles are from the same publisher. Afterwards, the click-through was calculated for each category. The results for both the baseline users and diverse users are presented in Figure 8.9. The statistical measures related to this data can be found in Table B.3 and Table B.4 in Appendix B.



(a) Baseline (b) Diverse

Figure 8.9: Mean and standard error of click-through rate for different levels of source diversity

Figure 8.9 illustrates clearly that no significant difference can be found in the click-through rate between two or three different publishers in the recommendation set for baseline users. Additionally, the results of the statistical measurement in Table B.3 and Table B.4 indicate that also for the diverse users, no significant difference was found.

ARTICLE LENGTH

As described in the experimental setup, it is assumed that the number of words of the recommendations can have an influence on the completion rate. Therefore, a statistical correlation analysis was conducted to assess whether both variables are related. Table 8.8 includes the Spearman's rank coefficient for both the baseline users and the diverse users. From these results, it can be seen that no significant correlation can be found between the completion rate and the number of words of a recommendation for neither the baseline and diverse users.

Group	ρ **statistic**	ρ p**-value**
Baseline	-0.26	0.12
Diverse	-0.19	0.27

Table 8.8: Spearman's rank correlation coefficient for correlation between the completion rate and the number of words of the recommendation

8.4. CONCLUSION

In addition to the offline evaluation, the proposed method was evaluation using an online study on the Blendle platform. For that purpose, Blendle has implemented a new functionality, such that three recommendations can be provided below the content of the editorial selected 'must reads' every day. Due to contractual limitations between Blendle and their publishers, recommendations could be provided for no more than two articles per day. During 12 days in a period of two weeks, two separated groups of 1038 users received recommendation from either the baseline method or diversified method. For that purpose, four different data sets of possible recommendations on specific ongoing and disputed topics were prepared. The reading behaviour of the two different groups were analysed using the click-through rate between the original article and the recommendation, the completion rate of the recommendation and the number of users that selected the recommendations as their favorite, also called the heart rate. Additional, the influence of multiple properties of the recommendations on the results were analysed, including the source diversity, presentation characteristics and article length. More details on the experimental setup can be found in section 8.2. Based on the results of the online evaluation, the section describes the conclusions that can be drawn related to the main research Question:

Main RQ: *How is reading behaviour affected by viewpoint diverse news recommendations and how they are presented?*

First, regarding the implicit and explicit measures of the reading behaviour, only significant results were found for the click-through rate per recommendation set. The results across all topics indicated a preference for the baseline recommendation sets: the click-through rate was 6.5% than for diverse recommendation sets. This was also found in the results per topic for the Corona recommendation sets. The other three measures of the reading behaviour, including the click-through rate per recommendation, the completion rate and the heart rate of a recommendation, also showed a slightly smaller mean for diverse user, compared to baseline users. However, none of these difference were found significant. Therefore, it can be concluded the the impact of the viewpoint diversification on the reading behaviour is small, according to the measures that were used in this evaluation.

Additionally, some conclusion can be made regarding the effect of multiple properties of the recommendations on the reading behaviour. First, the influence of the source diversity of the recommendation sets on the click-through rate calculated per set was assessed. Thereby, the source diversity was defined as the number of different publishers in a recommendation set of three. The results show that for both user groups each set contains either two or three different publishers but no significant differences on the influence of the property on the click-through rate can be found. Secondly, it was assumed that the article length, the number of words of an article, can have an influence on the completion rate of an article. However, for both user groups, no significant correlation was found between the article length and the completion rate. Finally, some interesting findings related to the presentation characteristics of the recommendations can be obtained from the results. As described before, a recommendation on the Blendle platform

is presented as rectangular box, including a title, publisher logo and a number of hearts, representing the number of users who selected the article as their favorite. Additionally, the box can be presented with a thumbnail image on the background and the title can be replaced by a custom editorial title. Regarding the title of the recommendation, no significant influence was found of both the inclusion of an editorial title and the number of words of the title on the click-through rate for both users groups. In contrast, both the inclusion of a thumbnail image and the number of presented hearts did have an impact on the click-through rate of diverse users. The click-through rate of recommendation with a thumbnail image was on average 3.1% higher for users in the received viewpoint diverse recommendations. Similarly, a modular positive correlation was found between the presented number of hearts and the click-through rate for diverse users. These results show that presentation characteristics indeed have a significant influence of the reading behaviour and thus, are important factors to analyse when addressing viewpoint diversification in online news media.

9

Discussion and Limitations

9.1. Introduction

In the previous two chapters the results of both the offline and online evaluation were presented. Based on these results, conclusions related to research questions were drawn. This chapter compares the findings with related research, discusses the implications of the results and presents the identified limitations. First, the findings of the offline evaluation are discussed. The findings related to the online discussion complete the chapter.

9.2. Offline Evaluation

According to the conclusion in section 7.4, the offline evaluation indicated that the proposed method is capable of increasing the viewpoint diversity of recommendation lists according the metric defined by Tintarev et al. [82]. Thereby, the average viewpoint diversity score across all topics increased from 0.55 to 0.79 for an increasing level of diversity in the MMR-algorithm. Simultaneously, the average relevance score decreased from 0.58 to 0.27. As described in the literature in section 2.2, the work by Tintarev et al. is, to our best knowledge, the only comparable work that proposed an algorithmic viewpoint diversification method [82]. During an offline evaluation on 386 articles from 17 different news outlets, a grid search was used to find the optimal model variables in terms of viewpoint diversity. The best performing model variables were found to be the initial, handcrafted setting yielding a viewpoint diversity score of 0.41. Remarkably, this score is considerably smaller than the maximum average value of 0.79 found in this work.

By taking a closer look at the viewpoint diversity metric, a possible factor that could have influenced this difference can be identified. As described in the experiment setup of the offline evaluation in section 7.2, the viewpoint diversity metric that is used in both studies is defined as the Intra-list diversity, which measures the average distance between every pair of articles. Distance, then, is implemented as 50% the distance between the extracted channels and 50% the extracted LDA topics. Where Tintarev et al. decided to exclude the LDA topic model from the diversification method to prevent any interference with the evaluation metric, the diversification method in this work still depends on

a LDA topic-model. Therefore, the difference in viewpoint diversity scores between the methods can possible be induced by interference of metadata between the viewpoint diversity metric and diversification method in this work.

Related to this, the offline evaluation could have benefit from a setup in which it was possible to assess the contribution of the individual framing aspects to the global viewpoint diversity score per article. Thereby, the relation between the extracted framing metadata and the viewpoint diversity score could be evaluated in more detail. For example, it could have been evaluated if specific aspects of the diversity function are responsible for the notable effects of higher levels of diversity on the publisher ratio. Moreover, a clear limitation of this work involves the minimum evaluation of how well the extracted metadata covers the framing function. Based on such analysis, a better comparison between different extraction methods can be made and individual extraction methods can be optimised. However, as described in the literature study in section 2.4, framing analysis still largely depends on human experts. Therefore, future research should address how framing can be defined, conceptualised and evaluated in the computational domain. This would, for example, enable the evaluation of the structure heuristic, the main result of the focus group on which the extraction pipeline highly depends.

Related to this, a broader discussion can be conducted about the viewpoint diversity metric that was used. As described in detail in the literature study in section 2.2, most approaches on diversification define diversity as the opposite of similarity and propose methods that are based on topic diversity. Diversity in news media, however, is understood as as multiperspectivity or a diversity of viewpoints. Therefore, this work aimed to bridge viewpoint diversity in the social domain to the computer science domain. However, besides the need for novel diversification models based on viewpoint diversity, it can be argued that the demand for a suitable viewpoint metric is even larger. Mainly, because the relation between the metric and the way viewpoint diversity is understood in literature is indistinct. The metric, defined the average distance between the channels and LDA-topics of pair of articles, appears to be more related to topic diversity than multiperspectivity, the way social sciences describes viewpoint diversity. The validity of the metric to evaluate viewpoint diversification can thus be questioned. Since the evaluation of novel diversification methods based on viewpoint diversity depends on an adequate metric, research on the development of such a metric should be a priority in this domain. Additionally, it may be worth the consider the role of source diversity in such a metric. As described in the literature study in section 2.3, two main approaches to assess viewpoint diversity can distinguished: methods based on content diversity and methods based on source diversity. Although approaches that use source diversity are more popular, scholars generally agree that viewpoint diversity can only by achieved by fostering content diversity. Because, multiple sources can still refer to the same point of view [89]. Based on these findings, this study used a content-based approach. From the results of the offline evaluation, it became clear that increasing levels of content diversity excludes multiple publishers and thus, decreases source diversity. Moreover, some specific publishers got amplified remarkably for high levels of content diversity. Therefore, viewpoint diversification methods could benefit from considering both content and source diversity.

Besides the viewpoint diversity metric, the offline study could have benefit from

a more sophisticated measure of the relevance between the recommendation and the original article. As described in section 4.5, due to the focus of the thesis on the viewpoint diversification method, the relevance score was based on a simple TF-IDF statistic, calculated within the topic. Therefore, it can be questioned to which extent this measures addresses relevance.

Additionally, some other limitations of the offline evaluation could be identified. First, the method was only analysed on four topics. However, certain results of the influence of the values of λ were found to be specific per data set, such as the gradient of the viewpoint diversity and relevance scores, the average number of words and the publisher ratio. An increased number of topics could potentially reveal other results that hold across topic, that were not identified in this setup. Similar, the average number of articles in the current evaluation could be pointed out as limited. Especially, when considering the balance of publishers in the data sets; In figure 5.3 in section 5.4, it can be seen that, although 15 publishers are represented in the data sets, the data sets mainly consists of data sets from the three most prominent publishers. For example, De Volkskrant has a share of more than 35% across all topics. Due to the limited number of articles and the unbalance in terms of publishers, the inclusion of a wide variety of perspectives on a topic can be challenged. The current major limitations involve the manual effort to both retrieve articles from Elasticsearch that match a certain topic and the manual check that has to be conducted afterwards. Moreover, the validity of these manual steps to assess whether an article belongs to a certain topic can be doubted. Finally, although different values of the cross-validation variable k and recommendation list size s were assessed, these variables were not included in the cross-validation due to time consideration. The quality of analysing the influence of these parameters on the results could be improved by using nested cross-validation.

9.3. ONLINE EVALUATION

As described in the conclusion of the online evaluation in section 7.4, no major influence of viewpoint diversification on the reading behaviour was found. Only the results of the click-through rate calculated per recommendation set indicated a significant difference between the baseline and diverse users of 6.5%. However, the results of the click-through rate calculated per recommendation indicated no significant difference between the two user groups. Likewise, the other two measurements of the reading behaviour, including both the completion rate of recommendations and the ratio of users who selected a recommendation as their favorite, showed no significant difference between baseline and diverse users.

In reflection on the motivation of this study, these results can be seen as positive. The proposed diversification for news media is capable of enhancing the viewpoint diversity of news recommendation, while maintaining comparable measures of the reading behaviour of users. The results thus suggest that recommender systems are capable of preserving the quality standards of multiperspectival in automatic online news environments. Thereby, situations of extreme low diversity, known is filter bubbles, can be prevented as well.

These results are in contrast with the most comparable work of Tintarev et al., who proposed a viewpoint diversification method based on the MMR-algorithm with linguis-

tic features, such as gravity, complexity and emotional tone [82]. During a user study, 15 participants were asked to make a forced choice between a recommendation from the diverse list and a recommendation from the baseline list, after reading an article on the same topic. In line with their hypothesis, it was found that 66% of the participants chose the baseline article, compared with 33% who chose the diverse article. However, some considerable differences with the online study that was conducted in this work can be indicated. First, the scale of experiment was much larger, including more than 2000 users. Additionally, users in this study were not asked to make a forced choice. Instead, an observation of the reading behaviour of both user groups was made without consciousness of users of participating in an experiment. It can be argued that setup of this study simulates situation in a more realistic way.

As described before, previous diversification in recommender systems mainly focused on topic diversity and thus were not directly applicable in the news domain in which diversity is defined as multiperspectivity. However, it can be argued that there is still considerable interference between the diversification methods. For example, comparable extraction techniques, such as LDA-models, diversification methods, including MMR, and diversity metrics, such as Intra-list diversity, were used. Therefore, insights from this research domain can still be interesting in the light of this study. Diversification gained significant awareness in research on recommender systems, mainly as a solution to the over-fitting problem many of these systems were struggling with [43]. Diversification, however, turned out to have a negative influence on the traditional accuracy-based evaluation metrics. This induced a shift to beyond-accuracy metrics for recommender systems, including user satisfaction with the recommendation. Afterwards, multiple studies indicated a positive effect of diversification on the user satisfaction with the recommendations [94, 40, 90, 51, 19]. Additionally, research in this domain showed that the level of diversity can have an impact on the user satisfaction. Ziegler et al. showed that the user satisfaction peaks at 40% diversity. Moreover, it was suggested that different personalities have different preferences for properties, such as diversity, of recommendation lists. For example, Nguyen et al. show that the user satisfaction for an increasing level of diversity in a recommendation list remains stable for high introversion people, but decreases for low introversion people [60]. Based on these insights, some scholars proposed diversification methods tailored to the individual needs [91]. As described in the experimental setup in section 8.2.5, to ensure significant results were obtained during the online evaluation, it was decided to use $\lambda = 0$ which yielded a maximum viewpoint diversity score for all topics during the offline evaluation. Due to time consideration, no other values were evaluated. These findings in this related research domain, however, suggest that different levels of diversity can have a significant influence on the interaction of users with the system. Thus, future research should assess different levels of diversity in the online evaluation. Thereby, personalised levels of diversity could also be considered.

Related to the potential interference with other diversification approaches, a broader discussion can be carried out about the metrics that were measured in this study. In case such a comprehensive analysis of the influence of different vales for λ will be done, it will also be interesting to assess the interference with other diversification approaches in more detail. For example, the effect of the viewpoint diversification on the metrics

that are used in these studies, such as topic diversity, novelty and serendipity, can be measured. This could provide more insight on how viewpoint diversification methods are different from traditional diversification methods in recommender systems.

Additionally, the results shed light on the importance of how a recommendation is presented. As can be obtained in the results of the online study in section 8.3, multiple presentation properties, such as the inclusion of a thumbnail image, were shown to have a significant influence on the click-through rate of recommendations. Future research, thus, should not only address the capability of a model to enhance viewpoint diversity according to an offline metric, but also evaluate what presentation characteristics could impact the user's willingness to read multiperspectival news. Related research on *viewpoint-aware-interfaces*, which aim to explain the recommendation choices to users, can be seen as very valuable [81, 54].

Besides, some other limitations of the online evaluation could be identified. First, the variety of topics, for which the reading behaviour could be analysed, is very low. Although data sets on four different topics were prepared, the topic of corona gained so much attention during the time of the experiment that other topics faded into the background. Moreover, for the topic of the Black Lives Mater Movement, no recommendations could be provided. As a result, 18 of the 24 articles for which recommendations were provided addressed the topic of corona. Although higher completion rate of the baseline method was found to be significant for all topics, the generality of the results across topic can thus be questioned. If the experiment would be conducted during a longer period of time, the reading behaviour could be measured for a wider variety of topics. Additionally, an extended online evaluation could strengthen the evidence that viewpoint diversification only slightly affects the reading behaviour. Secondly, additional to different values for λ, other values for the list size s of the recommendation list could have been valuable to analyse. As described in the experimental setup in section 8.2, the current set size was fixed to three by Blendle. However, during the offline evaluation different effects of the lists size were obtained. For example for higher levels of diversity, the effects on the publisher ratio, such as the decrease of source diversity, became smaller for larger recommendation sets. Moreover, some related studies show the significant influence of the set size on the user satisfaction and choice difficulty, thereby supporting the need to analyse different set sizes for the recommendations [90]. Also, it must be noted that only users who frequently used the regular read further section below article content where selected for the experiment. Therefore, the click-through rates that were presented in this study are not representative for all Blendle users. Finally, the current setup of the online evaluation required a manual assessment to match the daily editorial 'must reads' with a data set on a certain topic. Although the topic was described before the start of experiment including some specific keywords, the manual check still relied on the context and knowledge of the researcher. The validity of this manual step can thus be questioned.

10

Conclusion and Future Work

10.1. Conclusion

Most studies on diversification define diversity as the opposite of similarity and propose methods that are based on topic diversity. Diversity in news media, however, is understood as multiperspectivity or a diversity of viewpoints and scholars generally agree that fostering diversity is the key responsibility of the press in a democratic society. Therefore, novel diversification methods are needed that are capable of enhancing the diversity of viewpoints in recommendation lists. In the end, however, online news readers should also be willing to consume viewpoint diverse news recommendation. Therefore, to enable true multiperspectivity in the online news environment, research should also address how the reading behaviour is affected by viewpoint diverse news recommendations and how they are presented. Therefore, the main research question of this research is defined as follows:

Main RQ: *How is reading behaviour affected by viewpoint diverse news recommendations and how they are presented?*

To be able to answer this question, four sub-questions were defined that needed to be answered preliminary.

1. *How is diversity defined in the context of news media? What conceptualisation can be used to diversify news recommendation?*

   To answer this research question, a literature study was conducted on the diversity in news media and framing theory. Diversity in news media is understood as multiperspectivity or a diversity of viewpoints. Two main approaches to assess viewpoint diversity can be distinguished: methods based on content diversity and methods based on source diversity. Scholars generally agree that viewpoint diversity can only by achieved by fostering content diversity [89]. Among studies

on content diversity, frames are generally seen as one of the most suitable conceptualisations of this concept. The definition of a frame including four framing aspects by Entman is most commonly used [20]. This definition states that framing includes the selection of "some aspects of perceived reality and make them more salient in a communicating text, in such a way as to promote a particular definition of a problem, causal interpretation, moral evaluation and/or treatment recommendation for the item described" [20]. Similar to the work of Matthes and Kohring, this research conceptualised diversity as the diversity of these four framing aspects.

2. *What metadata can be related to this conceptualisation and which methods and tools can be used for the extraction of this data?*

   Based on the results of a focus group with three experts in the field of journalism, communication or news media, it was decided to focus on articles of the type background analysis and opinion pieces. In these types of articles, information related to certain framing aspects can often be found at specific places in the article. The introductory paragraphs are often used to describe the main problem under investigation. This can be related to the first framing aspect. Secondly, in the body of the article, different actors that contribute to the main issue under consideration are discussed, together with an evaluation of these forces. This can be related to the second and third framing aspect. Finally, the concluding paragraphs can be used by the author to provide any suggestions to improve or solve the problem. This can be related to fourth framing function. This structure was used as main heuristic in retrieving metadata related to each framing aspect as described by Entman. During the literature study, most relevant natural language processing techniques were discussed. Due to time limitations, one possible extraction pipeline was implemented and optimised using a setup of multiple NLP-toolkits. A justification of the choices that were made during the implementation of this pipeline can be found in chapter 4.

3. *How can this metadata be combined to a measure for viewpoint diversity that can be used in a recommender system?*

   The Maximal Marginal Relevance (MMR) algorithm is assumed to be most suitable for the diversification method. Thereby, a similar approach is taken as the most comparable work by Tintarev et al. [82]. For that purpose, a distance function related to each framing function was needed. Such a function measures the diversity of two articles in terms of the framing aspect. In the methodology in chapter 4, the choice of the distance function for each framing function is justified. The total diversity measure was implemented as the weighted linear combination of the four framing functions.

4. *Is the proposed method capable of increasing the viewpoint diversity of recommendation lists, according to a metric from literature?*

During an offline evaluation, the performance of different setups of the diversification method were evaluated according to a viewpoint diversity metric from literature, against a baseline method that was fully based on relevance [82]. For that purpose, four data sets on specific topics were created, each including between 40 and 70 different articles. In a cross-validation setup, the model parameters were optimised for each data set by means of a grid search. Additionally, the influence of different levels of viewpoint diversity, cross-fold setting and size of the recommendation size on multiple aspects of the recommendation list were analysed, including the viewpoint diversity score, relevance, rank correlation, average number of words and publisher ratio.

Based on the results, it was concluded that the method is capable of increasing the viewpoint diversity of recommendations lists, according to the metric defined by Tintarev et al. [82]. For decreasing values of λ, which implies a larger contribution of diversity in the MMR-algorithm, the viewpoint diversity scores increase among all evaluated topics, values for the size of the recommendation list s and values for the cross-validation parameter k. Among some additional results, the influence of the level of diversity on the publisher ration was found most remarkable. Increased levels of diversity result in the exclusion of certain publishers for all topics. Also, for the majority of the topics some publishers get amplified remarkably.

Finally, related to the main research question, an online study was conducted. For that purpose, Blendle implemented a new functionality on their platform, enabling the presentation of three recommendations below the content of an article on the same topic. During a two-week experiment including 2076 users, recommendations were provided below 24 articles on three different topics. Thereby, half of the users received recommendations from the baseline method, fully based on relevance, while the other users received viewpoint diverse recommendation, based on the best performing parameters in terms of viewpoint diversity score according to the offline evaluation. Three different aspects of the reading behaviour were observed, including the click-through rate from the original article to the recommendation calculated per recommendation and per recommendation set, the completion rate of a recommendation and the heart rate, the number of users who selected the recommendation as favorite. Additionally, among some other properties, the influence of the presentation characteristics of a recommendation on the click-through rate was evaluated.

Generally, no major differences were found in the reading behaviour of both user groups. Only the results of the click-through rate calculated per recommendation set indicated a significant difference of 6.5% to the advantage of the baseline users. However, the results of the click-through rate calculated per recommendation, the completion rate of recommendations and the ratio of users who selected a recommendation as their favorite, showed no significant difference between baseline and diverse users. Interesting enough, the results regarding the influence of the presentation characteristics on the click-through indicated some significant differences. For the diverse users the inclusion of a thumbnail image has a positive effect on the click-through, increasing 3.1%. Similarly, the number of hearts, representing the number of users that selected the article as the their favorite, was found to be positively correlated with the click-through rate for

diverse users. Therefore, these results suggest the future research on how recommendation can be presented is just as important as novel viewpoint diversification methods to truly achieve multiperspectivity in automated online news environments.

10.2. FUTURE WORK

Based on the discussion in chapter 9, multiple suggestions for future research on this topic can be made. Below, some specific recommendations are presented:

- **Presentation Characteristics**
  One of the major findings of the research include the significant influence of the presentation characteristics of news recommendations on the reading behaviour. Based on these results, it can be argued that future research should analyse this relationship in more detail. For example, the assessment of presentation characteristics in this study was bounded to how articles are presented on the Blendle platform. A more general approach can be chosen to analyse the contributions of these properties more comprehensively. Additionally, extended evaluation can be performed including an increasing number of users and recommendation. Based on the results of the online evaluation, it can be argued the future research on how recommendation can be presented is just as important as novel viewpoint diversification methods to truly achieve multiperspectivity in automated online news environments.

- **Viewpoint Diversity Metric**
  As described in the discussion of the offline evaluation in section 9.2, the validity of the viewpoint diversity metric that was used in this work can be questioned. The current metric seems more related to topic diversity, the way diversity is generally understood in the research domain of recommender systems, than as multiperspectivity, the way diversity is understood in the context of news media. Since the evaluation of novel diversification methods based on viewpoint diversity depends on an adequate metric, research on the development of such a metric should be a priority in this domain. As described in the results of the offline evaluation in section 7.3, the proposed content diversification method was found to have a considerable effect on the source diversity. Therefore, future research on the viewpoint diversity metric should address how aspects relate to each other.

- **Additional Types of Content**
  The current method focused on the extraction of metadata from textual content of the article. However, multiple other forms of content can be used. The most obvious example involves the inclusion of visual content, such as images in the analysis. Additionally, from the focus group session, it became clear that contextual information about a topic can also be essential to reveal a certain frame.

- **Evaluation of Extraction Methods**
  As described in the discussion of the offline evaluation in section 9.2, a clear limitation of this work involves the minimum evaluation of how well the extracted metadata covers the framing function. For that purpose, future research should

address how framing can be defined, conceptualised and evaluated in the computational domain. For example, the current extraction methods highly depended on the structure of two common types of articles. Based on suitable evaluation methods, the validity of this approach could have been assessed. Also, it would enable the comparison between different extraction pipelines and the optimisation of specific models to extract framing data.

BIBLIOGRAPHY

[1] Maud L Adriaansen, Philip Van Praag, and Claes H De Vreese. "Substance matters: How news content can reduce political cynicism". In: *International Journal of Public Opinion Research* 22.4 (2010), pp. 433–457.

[2] Xavier Amatriain and Justin Basilico. "Netflix recommendations: beyond the 5 stars (part 1)". In: *Netflix Tech Blog* 6 (2012).

[3] Kenneth T Andrews and Neal Caren. "Making the news: Movement organizations, media attention, and the public agenda". In: *American sociological review* 75.6 (2010), pp. 841–866.

[4] Christian Baden and Nina Springer. "Com (ple) menting the news on the financial crisis: The contribution of news users' commentary to the diversity of viewpoints in the public debate". In: *European journal of communication* 29.5 (2014), pp. 529–548.

[5] Christian Baden and Nina Springer. "Conceptualizing viewpoint diversity in news discourse". In: *Journalism* 18.2 (2017), pp. 176–194.

[6] C Edwin Baker. *Media, markets, and democracy*. Cambridge University Press, 2001.

[7] Eytan Bakshy, Solomon Messing, and Lada A Adamic. "Exposure to ideologically diverse news and opinion on Facebook". In: *Science* 348.6239 (2015), pp. 1130–1132.

[8] Gregory Bateson. "A theory of play and fantasy; a report on theoretical aspects of the project of study of the role of the paradoxes of abstraction in communication." In: *Psychiatric research reports* 2 (1955), pp. 39–51.

[9] Slobodan Beliga, Ana Meštrović, and Sanda Martinčić-Ipšić. "An overview of graph-based keyword extraction methods and approaches". In: *Journal of information and organizational sciences* 39.1 (2015), pp. 1–20.

[10] W Lance Bennett. "An introduction to journalism norms and representations of politics". In: (1996).

[11] Rodney Benson. "What makes news more multiperspectival? A field analysis". In: *Poetics* 37.5-6 (2009), pp. 402–418.

[12] David M Blei, Andrew Y Ng, and Michael I Jordan. "Latent dirichlet allocation". In: *Journal of machine Learning research* 3.Jan (2003), pp. 993–1022.

[13] Jason Brownlee. *How to Calculate the KL Divergence for Machine Learning*. Oct. 2019. URL: https://machinelearningmastery.com/divergence-between-probability-distributions/.

[14] Björn Burscher et al. “Teaching the computer to code frames in news: Comparing two supervised machine learning approaches to frame analysis”. In: *Communication Methods and Measures* 8.3 (2014), pp. 190–206.

[15] Jaime Carbonell and Jade Goldstein. “The use of MMR, diversity-based reranking for reordering documents and producing summaries”. In: *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. 1998, pp. 335–336.

[16] Jihyang Choi. “Diversity in foreign news in US newspapers before and after the invasion of Iraq”. In: *International Communication Gazette* 71.6 (2009), pp. 525–542.

[17] Claes H De Vreese. “News framing: Theory and typology.” In: *Information design journal & document design* 13.1 (2005).

[18] Maunendra Sankar Desarkar and Neha Shinde. “Diversification in news recommendation for privacy concerned users”. In: *2014 International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE. 2014, pp. 135–141.

[19] Michael D Ekstrand et al. “User perception of differences in recommender algorithms”. In: *Proceedings of the 8th ACM Conference on Recommender systems*. ACM. 2014, pp. 161–168.

[20] Robert M Entman. “Framing: Toward clarification of a fractured paradigm”. In: *Journal of communication* 43.4 (1993), pp. 51–58.

[21] Robert M Entman and Steven S Wildman. “Reconciling economic and non-economic perspectives on media policy: Transcending the “marketplace of ideas””. In: *Journal of Communication* 42.1 (1992), pp. 5–19.

[22] Myra Marx Ferree et al. *Shaping abortion discourse: Democracy and the public sphere in Germany and the United States*. Cambridge University Press, 2002.

[23] William A Gamson and Andre Modigliani. “Media discourse and public opinion on nuclear power: A constructionist approach”. In: *American journal of sociology* 95.1 (1989), pp. 1–37.

[24] Herbert J Gans. *Democracy and the News*. Oxford University Press, 2003.

[25] Daniel Gildea and Daniel Jurafsky. “Automatic labeling of semantic roles”. In: *Computational linguistics* 28.3 (2002), pp. 245–288.

[26] Todd Giltin. *The whole world is watching: Mass media in the making and unmaking of the new left*. McGraw-Hill, 1980.

[27] Andrew B Goldberg et al. “May all your wishes come true: A study of wishes and how to recognize them”. In: *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. 2009, pp. 263–271.

[28] Robert A Hackett. “A hierarchy of access: Aspects of source bias in Canadian TV news”. In: *Journalism Quarterly* 62.2 (1985), pp. 256–277.

[29] Mario Haim, Andreas Graefe, and Hans-Bernd Brosius. "Burst of the filter bubble? Effects of personalization on the diversity of Google News". In: *Digital journalism* 6.3 (2018), pp. 330–343.

[30] Luheng He et al. "Deep semantic role labeling: What works and what's next". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2017, pp. 473–483.

[31] James K Hertog and Douglas M McLeod. "A multiperspectival approach to framing analysis: A field guide". In: *Framing public life*. Routledge, 2001, pp. 157–178.

[32] Daniel E Ho and Kevin M Quinn. "Viewpoint diversity and media consolidation: An empirical study". In: *Stan. L. Rev.* 61 (2008), p. 781.

[33] Brett Hutchins and Libby Lester. "Environmental protest and tap-dancing with the media in the information age". In: *Media, Culture & Society* 28.3 (2006), pp. 433–451.

[34] Paul Jaccard. "Étude comparative de la distribution florale dans une portion des Alpes et des Jura". In: *Bull Soc Vaudoise Sci Nat* 37 (1901), pp. 547–579.

[35] Hamed Jelodar et al. "Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey". In: *Multimedia Tools and Applications* 78.11 (2019), pp. 15169–15211.

[36] D Kahneman and A Tversky. "Choices, values, and frames". In: *The American psychologist* 39.4 (1984), pp. 341–350.

[37] Suvarna G Kanakaraddi and Suvarna S Nandyal. "Survey on parts of speech tagger techniques". In: *2018 International Conference on Current Trends towards Converging Technologies (ICCTCT)*. IEEE. 2018, pp. 1–6.

[38] Mozhgan Karimi, Dietmar Jannach, and Michael Jugovac. "News recommender systems–Survey and roads ahead". In: *Information Processing & Management* 54.6 (2018), pp. 1203–1227.

[39] Maurice George Kendall. "Rank correlation methods." In: (1948).

[40] Bart P Knijnenburg et al. "Explaining the user experience of recommender systems". In: *User Modeling and User-Adapted Interaction* 22.4-5 (2012), pp. 441–504.

[41] Klaus Krippendorff. "Estimating the reliability, systematic error and random error of interval data". In: *Educational and Psychological Measurement* 30.1 (1970), pp. 61–70.

[42] Anne C Kroon et al. "Victims or perpetrators? Explaining media framing of Roma across Europe". In: *European Journal of Communication* 31.4 (2016), pp. 375–392.

[43] Matevž Kunaver and Tomaž Požrl. "Diversity in recommender systems–A survey". In: *Knowledge-Based Systems* 123 (2017), pp. 154–162.

[44] Christopher D Manning et al. "The Stanford CoreNLP natural language processing toolkit". In: *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*. 2014, pp. 55–60.

[45] Diego Marcheggiani and Ivan Titov. “Encoding sentences with graph convolutional networks for semantic role labeling”. In: *arXiv preprint arXiv:1703.04826* (2017).

[46] Matt Marshall. “Aggregate Knowledge raises $5 M from Kleiner, on a roll”. In: *Venture Beat* December 10 (2006).

[47] Shannon Rossi Martin. “Proximity of event as factor in selection of news sources”. In: *Journalism Quarterly* 65.4 (1988), pp. 986–989.

[48] Andrea Masini et al. “Measuring and explaining the diversity of voices and viewpoints in the news: A comparative study on the determinants of content diversity of immigration news”. In: *Journalism Studies* 19.15 (2018), pp. 2324–2343.

[49] Jörg Matthes. “What’s in a frame? A content analysis of media framing studies in the world’s leading communication journals, 1990-2005”. In: *Journalism & Mass Communication Quarterly* 86.2 (2009), pp. 349–367.

[50] Jörg Matthes and Matthias Kohring. “The content analysis of media frames: Toward improving reliability and validity”. In: *Journal of communication* 58.2 (2008), pp. 258–279.

[51] Denis McQuail. *Media performance: Mass communication and the public interest*. Vol. 144. Sage London, 1992.

[52] Walaa Medhat, Ahmed Hassan, and Hoda Korashy. “Sentiment analysis algorithms and applications: A survey”. In: *Ain Shams engineering journal* 5.4 (2014), pp. 1093–1113.

[53] Samaneh Moghaddam. “Beyond sentiment analysis: mining defects and improvements from customer feedback”. In: *European conference on information retrieval*. Springer. 2015, pp. 400–410.

[54] Sayooran Nagulendra and Julita Vassileva. “Understanding and controlling the filter bubble through interactive visualization: a user study”. In: *Proceedings of the 25th ACM conference on Hypertext and social media*. 2014, pp. 107–115.

[55] Philip M Napoli. “Deconstructing the diversity principle”. In: *Journal of communication* 49.4 (1999), pp. 7–34.

[56] Sapna Negi. “Suggestion mining from text”. PhD thesis. NUI Galway, 2019.

[57] Sapna Negi et al. “A study of suggestions in opinionated texts and their automatic detection”. In: *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*. 2016, pp. 170–178.

[58] Nic Newman, David AL Levy, and Rasmus Kleis Nielsen. *Reuters Institute digital news report 2015: Tracking the future of news*. Reuters Institute for the Study of Journalism, 2015.

[59] Tien T Nguyen et al. “Exploring the filter bubble: the effect of using recommender systems on content diversity”. In: *Proceedings of the 23rd international conference on World wide web*. ACM. 2014, pp. 677–686.

[60] Tien T Nguyen et al. “User personality and user satisfaction with recommender systems”. In: *Information Systems Frontiers* 20.6 (2018), pp. 1173–1189.

[61] Eli Pariser. *The filter bubble: How the new personalized web is changing what we read and how we think*. Penguin, 2011.

[62] Alexandre Pinto, Hugo Gonçalo Oliveira, and Ana Oliveira Alves. "Comparing the performance of different NLP toolkits in formal and social media text". In: *5th Symposium on Languages, Applications and Technologies (SLATE'16)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik. 2016.

[63] Mauro P Porto. "Frame diversity and citizen competence: Towards a critical approach to news quality". In: *Critical Studies in Media Communication* 24.4 (2007), pp. 303–321.

[64] Mauro P Porto. "Framing controversies: television and the 2002 presidential election in Brazil". In: *Political Communication* 24.1 (2007), pp. 19–36.

[65] Mauro P Porto. "Media framing and citizen competence: Television and audiences' interpretations of politics in Brazil." In: (2002).

[66] Janardhanan Ramanand, Krishna Bhavsar, and Niranjan Pedanekar. "Wishful thinking-finding suggestions and'buy'wishes from product reviews". In: *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*. 2010, pp. 54–61.

[67] Adam D Rennhoff and Kenneth C Wilbur. "Market-based measures of viewpoint diversity". In: *Information Economics and Policy* 26 (2014), pp. 1–11.

[68] Paul Resnick and Hal R Varian. "Recommender systems". In: *Communications of the ACM* 40.3 (1997), pp. 56–59.

[69] Charlotte Ryan, Kevin M Carragee, and Cassie Schwerner. "Media, movements, and the quest for social justice". In: (1998).

[70] Gerard Salton and Michael J McGill. *Introduction to modern information retrieval*. mcgraw-hill, 1983.

[71] Dipanjan (DJ) Sarkar. *A Practitioner's Guide to Natural Language Processing (Part I) — Processing & Understanding Text*. Dec. 2018. URL: https://towardsdatascience.com/a-practitioners-guide-to-natural-language-processing-part-i-processing-understanding-text-9f4abfd13e72.

[72] J Ben Schafer, Joseph A Konstan, and John Riedl. "E-commerce recommendation applications". In: *Data mining and knowledge discovery* 5.1-2 (2001), pp. 115–153.

[73] Dietram A Scheufele. "Framing as a theory of media effects". In: *Journal of communication* 49.1 (1999), pp. 103–122.

[74] Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. *Introduction to information retrieval*. Vol. 39. Cambridge University Press Cambridge, 2008.

[75] Dhavan V Shah et al. "News framing and cueing of issue regimes: Explaining Clinton's public approval in spite of scandal". In: *Public Opinion Quarterly* 66.3 (2002), pp. 339–370.

[76] Jackie Smith et al. "From protest to agenda building: Description bias in media coverage of protest events in Washington, DC". In: *Social Forces* 79.4 (2001), pp. 1397–1423.

[77] Paul M Sniderman and Sean M Theriault. “The structure of political argument and the logic of issue framing”. In: *Studies in public opinion: Attitudes, nonattitudes, measurement error, and change* (2004), pp. 133–65.

[78] David A Snow et al. “Frame alignment processes, micromobilization, and movement participation”. In: *American sociological review* (1986), pp. 464–481.

[79] Jesper Strömbäck. “In search of a standard: Four models of democracy and their normative implications for journalism”. In: *Journalism studies* 6.3 (2005), pp. 331–345.

[80] Zhixing Tan et al. “Deep semantic role labeling with self-attention”. In: *Thirty-Second AAAI Conference on Artificial Intelligence*. 2018.

[81] Nava Tintarev. “Presenting diversity aware recommendations: Making challenging news acceptable”. In: (2017).

[82] Nava Tintarev et al. “Same, same, but different: algorithmic diversification of viewpoints in news”. In: *Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization*. 2018, pp. 7–13.

[83] Jan Van Cuilenburg. “On competition, access and diversity in media, old and new: Some remarks for communications policy in the information age”. In: *New media & society* 1.2 (1999), pp. 183–207.

[84] Baldwin Van Gorp. “The constructionist approach to framing: Bringing culture back in”. In: *Journal of communication* 57.1 (2007), pp. 60–78.

[85] Elisabeth A Van Zoonen. “The women’s movement and the media: Constructing a public identity”. In: *European Journal of Communication* 7.4 (1992), pp. 453–476.

[86] Saúl Vargas and Pablo Castells. “Rank and relevance in novelty and diversity metrics for recommender systems”. In: *Proceedings of the fifth ACM conference on Recommender systems*. ACM. 2011, pp. 109–116.

[87] Amar Viswanathan et al. “Suggestion Mining from Customer Reviews.” In: *AMCIS*. 2011.

[88] Rens Vliegenthart. “Framing in mass communication research–An overview and assessment”. In: *Sociology Compass* 6.12 (2012), pp. 937–948.

[89] Paul S Voakes et al. “Diversity in the news: A conceptual and methodological framework”. In: *Journalism & Mass Communication Quarterly* 73.3 (1996), pp. 582–593.

[90] Martijn C Willemsen, Mark P Graus, and Bart P Knijnenburg. “Understanding the role of latent feature diversification on choice difficulty and satisfaction”. In: *User Modeling and User-Adapted Interaction* 26.4 (2016), pp. 347–389.

[91] Wen Wu, Li Chen, and Yu Zhao. “Personalizing recommendation diversity based on user personality”. In: *User Modeling and User-Adapted Interaction* 28.3 (2018), pp. 237–276.

[92] Vikas Yadav and Steven Bethard. “A survey on recent advances in named entity recognition from deep learning models”. In: *arXiv preprint arXiv:1910.11470* (2019).

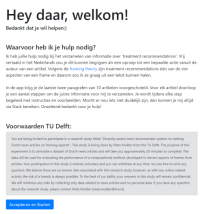
[93] Mi Zhang and Neil Hurley. “Avoiding monotony: improving the diversity of recommendation lists”. In: *Proceedings of the 2008 ACM conference on Recommender systems*. ACM. 2008, pp. 123–130.

[94] Cai-Nicolas Ziegler et al. “Improving recommendation lists through topic diversification”. In: *Proceedings of the 14th international conference on World Wide Web*. ACM. 2005, pp. 22–32.

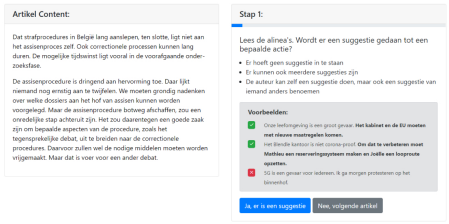
A

Crowdsourcing Web Application

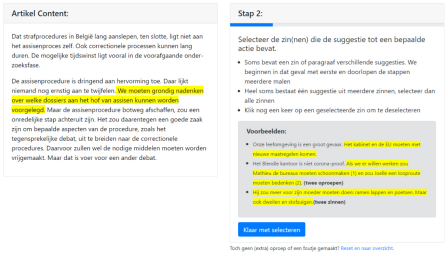
A



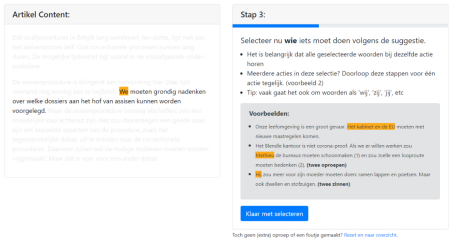
0. Read and accept terms and conditions TU Delft



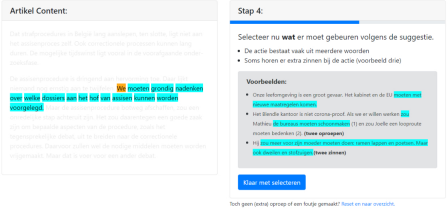
1. Is there any suggestion?



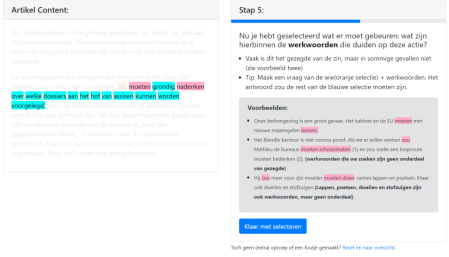
2. Select sentences of suggestion



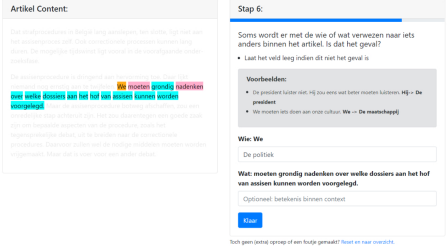
3. Select the subject of the suggestion



4. Select what is suggested



5. Select verbs in suggestion



6. Disambiguation

7. Overview

Figure A.1: Visual overview of crowdsource platform

B

Statistical Measures of Online Evaluation

B

Click-through per Recommendation

Topic	Group	Mean	Error	Shapiro-Wilk test
Total	Baseline	0.11	0.011	0.91, 0.0064
	Diverse	0.087	0.0083	0.92, 0.01
Corona	Baseline	0.11	0.012	0.96, 0.29
	Diverse	0.09	0.01	0.9, 0.021
U.S Elections	Baseline	0.13	0.044	0.72, 0.011
	Diverse	0.095	0.017	0.93, 0.62
Big Tech	Baseline	0.086	0.024	0.88, 0.3
	Diverse	0.062	0.022	0.85, 0.19

Click-through per Recommendation Set

Topic	Group	Mean	Error	Shapiro-Wilk test
Total	Baseline	0.31	0.016	0.96, 0.41
	Diverse	0.25	0.016	0.95, 0.27
Corona	Baseline	0.32	0.018	0.94, 0.32
	Diverse	0.25	0.018	0.94, 0.32
U.S Elections	Baseline	0.31	0.025	0.84, 0.19
	Diverse	0.28	0.041	0.69, 0.01
Big Tech	Baseline	0.25	0.11	/
	Diverse	0.18	0.035	/

Completion Rate

Topic	Group	Mean	Error	Shapiro-Wilk test
Total	Baseline	0.43	0.026	0.97, 0.36
	Diverse	0.39	0.029	0.98, 0.81
Corona	Baseline	0.43	0.033	0.96, 0.31
	Diverse	0.41	0.034	0.96, 0.36
U.S Elections	Baseline	0.42	0.055	0.98, 0.96
	Diverse	0.39	0.059	0.96, 0.83
Big Tech	Baseline	0.42	0.075	0.92, 0.53
	Diverse	0.3	0.11	0.91, 0.47

Heart Ratio

Topic	Group	Mean	Error	Shapiro-Wilk test
Total	Baseline	0.22	0.038	0.78, 4.6e-06
	Diverse	0.23	0.083	0.4, 6e-11
Corona	Baseline	0.26	0.049	0.78, 5.9e-05
	Diverse	0.29	0.12	0.43, 8.5e-09
U.S Elections	Baseline	0.16	0.05	0.96, 0.79
	Diverse	0.099	0.033	0.81, 0.072
Big Tech	Baseline	0.097	0.073	0.72, 0.016
	Diverse	0.08	0.08	0.55, 0.00013

Influence of Presentation Characteristics

Topic	Group	Mean	Error	Shapiro-Wilk test
Thumbnail	With	0.11	0.0085	0.94, 0.032
	Without	0.088	0.011	0.83, 6.1e-05
Editorial Title	With	0.087	0.0058	0.94, 0.57
	Without	0.099	0.0079	0.91, 0.00023

Table B.1: Mean, error and Shapiro-Wilk test for results per topic of the click-through rate per recommendation, the click-through rate per recommendation set, the completion rate and the heart rate

Per Recommendation

Topic	Levene's test	Student t-test	Welch's t-test	Mann-Whitney U test
Total	2.2, 0.14	1.5, 0.13	1.5, 0.13	570.0, 0.1
Corona	1.4, 0.24	1.2, 0.25	1.2, 0.24	280.0, 0.13
U.S. Elections	0.52, 0.49	0.63, 0.54	0.63, 0.55	17.0, 0.47
Big Tech	0.061, 0.81	0.71, 0.5	0.71, 0.5	8.0, 0.2

Per Recommendation Set

Topic	Levene's test	Student t-test	Welch's t-test	Mann-Whitney U test
Total	0.048, 0.83	2.9, 0.0054	2.9, 0.0054	160.0, 0.004
Corona	0.52, 0.48	2.9, 0.0063	2.9, 0.0063	80.0, 0.005
U.S. Elections	0.11, 0.75	0.53, 0.61	0.53, 0.62	7.0, 0.44
Big Tech	/	0.59, 0.61	0.59, 0.61	2.0, 0.35

Completion Rate

Topic	Levene's test	Student t-test	Welch's t-test	Mann-Whitney U test
Total	0.56, 0.46	0.94, 0.35	0.94, 0.35	600.0, 0.17
Corona	0.008, 0.93	0.44, 0.66	0.44, 0.66	320.0, 0.35
U.S. Elections	0.19, 0.67	0.43, 0.68	0.43, 0.68	16.0, 0.41
Big Tech	0.06, 0.81	0.9, 0.39	0.9, 0.4	7.0, 0.15

Heart Rate

Topic	Levene's test	Student t-test	Welch's t-test	Mann-Whitney U test
Total	0.43, 0.51	-0.072, 0.94	-0.071, 0.94	580.0, 0.14
Corona	0.58, 0.45	-0.24, 0.81	-0.23, 0.82	300.0, 0.25
U.S. Elections	1.5, 0.25	0.95, 0.37	0.95, 0.37	13.0, 0.23
Big Tech	0.025, 0.88	0.16, 0.88	0.16, 0.88	11.0, 0.4

Influence of Presentation Characteristics

Property	Levene's test	Student t-test	Welch's t-test	Mann-Whitney U test
Thumbnail	0.42, 0.52	1.4, 0.18	1.4, 0.18	490.0, 0.02
Editorial Title	5.9, 0.017	-0.56, 0.58	-1.3, 0.2	260.0, 0.48

Table B.2: Overview of the Levene's test, Student t-tes, Welch's t-test and Mann-Whitney U test for results per topic of the click-through rate per recommendation, the click-through rate per recommendation set, the completion rate and the heart rate

Property	Variation	Mean	Error	Shapiro-Wilk test
Sources: 2	Baseline	0.32	0.022	0.96, 0.69
	Diverse	0.23	0.022	0.92, 0.18
Sources: 3	Baseline	0.31	0.025	0.95, 0.64
	Diverse	0.28	0.017	0.93, 0.46
With Thumbnail	Baseline	0.11	0.013	0.97, 0.76
	Diverse	0.1	0.011	0.85, 0.0059
Without Thumbnail	Baseline	0.1	0.019	0.84, 0.005
	Diverse	0.071	0.011	0.85, 0.011
With Editorial Title	Baseline	0.085	0.012	0.9, 0.43
	Diverse	0.089	0.0033	0.85, 0.23
Without Editorial Title	Baseline	0.11	0.012	0.92, 0.017
	Diverse	0.087	0.0094	0.9, 0.0077

Table B.3: Mean, error and Shapiro-Wilk test for results of click-through for different data properties

Property	Group	Levene's test	Student t-test	Welch's t-test	Mann-Whitney U test
Sources	Baseline	0.0058, 0.94	0.18, 0.86	0.18, 0.86	68.0, 0.43
	Diverse	2.6, 0.12	-1.5, 0.15	-1.7, 0.1	45.0, 0.095
Thumbnail	Baseline	0.64, 0.43	0.38, 0.7	0.38, 0.7	150.0, 0.16
	Diverse	0.014, 0.91	2.0, 0.056	2.0, 0.055	88.0, 0.01
Editorial Title	Baseline	2.8, 0.1	-0.72, 0.47	-1.5, 0.15	57.0, 0.31
	Diverse	3.9, 0.055	0.065, 0.95	0.18, 0.86	43.0, 0.15

Table B.4: Results of Levene's test, student t-test, Welch's test and Mann-Whitney test for different data properties